Contents lists available at ScienceDirect

Journal of Responsible Technology

journal homepage: www.sciencedirect.com/journal/journal-of-responsible-technology

Research Article

Normative conflict resolution through human–autonomous agent interaction

Beverley Townsend [a], Katie J. Parnell [b,*], Sinem Getir Yaman [a], Gabriel Nemirovsky [a], Radu Calinescu [a]

[a] *University of York, York, England, UK*
[b] *University of Southampton, Southampton, England UK*

ARTICLE INFO

Keywords:
Normative conflict
Autonomous agents
Decision-making
Choice
Defeasibility
Decision ladder
SRK levels

ABSTRACT

We have become increasingly reliant on the decision-making capabilities of autonomous agents. These decisions are often executed under non-ideal conditions, offer significant moral risk, and directly affect human well-being. Such decisions may involve the choice to optimise one value over another: promoting safety over human autonomy, or ensuring accuracy over fairness, for example. All too often decision-making of this kind requires a level of normative evaluation involving ethically defensible moral choices and value judgements, compromises, and trade-offs. Guided by normative principles such decisions inform the possible courses of action the agent may take and may even change a set of established actionable courses.

This paper seeks to map the decision-making processes in normative choice scenarios wherein autonomous agents are intrinsically linked to the decision process. A care-robot is used to illustrate how a normative choice - underpinned by normative principles - arises, where the agent must ‘choose’ an actionable path involving the administration of critical or non-critical medication. Critically, the choice is dependent upon the trade-off involving two normative principles: respect for human autonomy and the prevention of harm. An additional dimension is presented, that of the inclusion of the urgency of the medication to be administered, which further informs and changes the course of action to be followed.

We offer a means to map decision-making involving a normative choice within a decision ladder using stakeholder input, and, using defeasibility, we show how specification rules with defeaters can be written to operationalise such choice.

1. Introduction

Autonomous agents, with the capability to engage in personalised interaction with human-users are, or will, in the future, be required to make real-time decisions that involve ‘difficult’ ethical considerations and normative trade-offs (Mittelstadt et al., 2016; Wiegel & van den Berg, 2009). In executing a course of action, an autonomous agent may be called upon to select between two or more nontrivial normatively-relevant alternatives (or ‘options’) requiring of the agent to make decisions premised on an array of alternative, principled, reasoned and justifiable choices. Choices that may require, for instance, prioritising safety over respect for human autonomy, harm-prevention over privacy, or individual interests over collective interests. We currently lack methods and approaches to understand how these normative choices can be made in an ethical manner. Our paper aims to fill this gap with the application and development of a human decision making model to capture conflict resolution in human-autonomy relationships.

Exploring ‘machine ethics’ and the design and moral reasoning of autonomous agents is not new (Anderson et al., 2004; Anderson & Anderson, 2007; Anderson & Anderson, 2011; Anderson & Anderson, 2018; Anderson et al., 2018; Moor, 2006; Dodig Crnkovic & C¸ürüklü, 2012; Dennis et al., 2016; Jiang et al., 2021). Anderson & Anderson (Anderson et al., 2006) describe a system, for example, that generates rules for weighting the different prima facie obligations so that ethicists can articulate more general principles that would otherwise be hard to discern in human decision-making. Dennis et al. (Dennis et al., 2016; Cervantes et al., 2016) propose a formal theoretical verification framework for ethical plan selection that can be used to assist autonomous agents to make complex decisions by requiring the agent to choose to execute, to the best of its beliefs, an ethical plan. Cervantes et al. offer a

[*] Corresponding author.
E-mail address: k.parnell@soton.ac.uk (K.J. Parnell).

https://doi.org/10.1016/j.jrt.2025.100114

Available online 24 February 2025

computational model for ethical decision-making that takes into account preferences, good and bad past experiences, ethical rules, and current emotional state when making the most appropriate choice. More recently, Jiang et al. (Jiang et al., 2021) have demonstrated Delphi, a model to predict moral judgements from machines trained on descriptive ethical judgement data. This is large-scale data involving normative judgements collected from nuanced compositional and socially sensitive everyday situations presented in a commonsense norm bank. The difficulty is, however, not only in determining the potential and actual ethical impact of a decision made by a model, but in understanding and documenting the factors, features, and reasons that informed the decision making process and the ultimate outcome.

We have shown elsewhere (Townsend et al., 2022) how high-level principles can be refined to lower-level explicitly formulated operational rules or 'evaluative standards'. Guided by such high-level principles, lower-level programmable rules are written that inform an agent's executable course of action. These non-functional rules are used to augment and complement the specifications of the agent and set out what the agent ought to do in a given scenario. However, autonomous agents may be required to not only implement the legal, ethical, social, and cultural norms and expectations associated with their roles, but also to select between two 'conflicting' normative principles. A 'normative conflict' would include a moral conflict and speaks to a situation where two competing actions ought to, but cannot, be performed (Horty, 2012). An example would be where respecting human autonomy (by the agent obeying a user's instruction) comes into conflict, or is in tension, with the requirement to prevent harm to the user (as a consequence of not administering critical medication). This requires resolution by negotiated justifiable trade-off involving human deliberative evaluation in a process of collaborative engagement and reflection as described in Townsend et al. (Townsend et al., 2022). Our interest, in this paper, is in determining how such normative conflicts and choices arise and can be mapped, so that agents can be designed to implement decisions involving normative principles, which, in turn, can inform the execution of alternative tasks. Specification rules, hedged with 'defeaters', are thus written directing a specific choice outcome.

To illustrate our approach, we consider a hypothetical normative choice scenario confronted by a social or care robot ('carebot'). A carebot is a supportive robotic tool used to care for the health of the elderly, children, and those living with disabilities (Vallor, 2020). The carebot is typically deployed in the user's home (or at a care home) – either working alongside human caregivers or on its own – with the primary role of aiding a user, for instance, in providing routine care and support functions such as reminding a user to take their medication or in administering such medication. These carebots may also be a source of companionship and comfort to the user and are expected to engage in social interactions with the user, by communicating, listening, responding, and reacting and making certain normatively-relevant choices. Jevtic et al. (Jevtíc et al., 2018) describe the development of a similar carebot. It is a personalised agent with a wide range of physical characteristics and abilities that can perform assistive dressing functions in close physical interaction with users. Autonomous agents of this type demonstrate a degree of sociability and of emotional perception, such as, engaging in highlevel interactive dialogue, responsiveness, and gesturing, and using voice recognition. We provide an example in the case study of how, in the course of such an interface, an agent is required to execute a task premised on the resolution of two conflicting normative choices. We then show how alternative paths may be mapped using a cognitive work analysis, and how implementable rules can be written based on specific choice outcomes. As the agent, in our example, is semi-autonomous, we acknowledge the evolving perspectives related to algorithmic software development in autonomous agents in the fields of strategic planning software, business intelligence, uncertain reasoning, and automated decision making. Aligned with certain aspects of these approaches, we offer a new solution for the following common problems: (a) the identification and expression of conflicting goals; (b) the application of knowledge-based human behaviour and uncertain reasoning to support automated decision-making, and (c) the mapping, resolution, and achievement of these goals.

2. Theoretical framework

From plural principles comes the opportunity for principles to 'conflict'. 'Normative conflict' involves at least two options, underpinned by obligation or duty, and is described as 'competing' or in 'tension' in the sense that the pursuit of one option can resist or oppose another in a certain context (Townsend et al., 2022). Such 'tensions' inevitably arise as we bridge principles to practice. Value-conflict then is a position where only one interest or value can be upheld in a given case. These are instances when only a single option can be selected and must succeed in overriding the other, that is to say, one principle must be sacrificed in favour of another. This is a situation that William (Williams, 1981) sees not as pathological, but something necessarily involved in reconciling human values.

While future work can consider how autonomous agents might come to make moral choices (e.g., through machine learning), for clarity, and for the purposes of this paper, it is not the agent that makes the moral choice. A choice is made by the human stakeholders, and is merely executed by the agent as informed by a series of pre-written rules. A 'normative choice scenario' is a scenario where a decision-maker's choice and the subsequent course of action it gives rise to, contains a normative (or moral) dimension which has an impact (be it adverse or otherwise). These normative choices include, as a part of the deliberation, consideration of various legal, ethical, social, and cultural norms. These interrelated 'norms' are the 'fundamental principles that govern the issues of how we should live and what we morally ought to do' or that which would on the whole make things go well (Driver, 2005). We use the word 'normative' in this article in the sense that it relates to norms or standards, especially of behaviour and value. This, we believe, is wide enough to encompass moral standards of right and wrong behaviour. We, however, stop short of making moral claims regarding the choice, only we say that the claim or choice relates to a norm or standard which may or may not have a moral dimension.

We present normative choice mapping using decision ladders, with broader application to social science, and offer a position that says more about how the human decision-makers making normative choices behaves or what they prefer to do, rather than how they ought to behave or what they morally ought to do. Decision ladders are applied to break the decision process down into different levels of information processing, in accordance with psychological behaviour models (Rasmussen, 1983). In doing so, they present a structured approach to review the different elements that inform a decision and the process of assessing options to select the one most valid to the currently situation (Jenkins et al., 2010). Using this approach, we suggest that a choice is the culmination and expression of an act of accountable, deliberative evaluation that captures what the human agent actually selects or does in a given choice context. Thus, in a given choice scenario, we can say something about the actionable form of practical human reasoning expressed as a choice in selecting one alternative above another. We therefore present this paper with the following caveat: we do not make an assertion as to the 'correctness' of a moral choice or to suggest that a choice is 'good' or as instrumentally good (or, 'good for something' or 'good for someone'). Thus, we do not categorise alternatives or indeed choices as either 'right' or 'wrong', 'good' or not, or something in between. What it is for a decision to be good depends as much upon what the decision is about, as about what it is to be good for something or for someone. As there are criteria or standards independent of human choices, preferences, and attitudes that govern whether a choice is good or not, we do not here attribute goodness to a choice. Only, we suggest, that a human decision-maker within a context must act in accordance with what might be considered the actions of a 'reasonable and responsible person': to act virtuously (or to a common good), to deliberate about options before acting, to make choices that seem to be the best for all affected, and to concern themselves with how their actions may adversely affect others. (Young, 2011)

Moral judgement is complex and is informed by factors such as differences in the individuals' lives, values, norms, culture, and religion. When making a choice, we call on the stakeholders or human decision-makers to consider John Rawls' epistemological technique of envisaging themselves as within the 'original position' behind a 'veil of ignorance' (Rawls, 1971) – that is to say, that all individuals are similarly situated: no one is able to design principles to favour their particular condition, and no one knows their place, class, or social status in society (Rawls, 1971). Rawls proposed the Veil of Ignorance (or 'VoI') to identify fair principles as a means to govern society. One way to interpret Rawls's epistemological technique is as a game in which, 'a person would choose for the design of a society in which his enemy is to assign him his place,' and develop a strategy to maximise their own well-being accordingly.

Compared to individuals who know their position, those behind the Rawlsian veil are more likely, upon reflection, to choose principles and outcomes that prioritise the worstoff (Weidinger & McKee, 2022). The outcome is therefore prioritarian rather than maximisation - meaning rules prioritise benefiting the worst off. In these circumstances, decision-makers are called upon to evaluate principles and consider alternatives on the basis that they too might be the recipient of the decision making. (Rawls, 1971) Experimental data shows that the VoI helps promote reasoning towards fairness, when asked to justify their decision, compared to control groups - thus producing more pro-social attitudes (Weidinger & McKee, 2022). These benefits are despite any initial differences in risk-attitudes. In other words, appeals to fairness and other pro-social attitudes are not likely to be reducible to an aversion against the possibility of being the worst off. Using the VoI as a procedural rule in developing the governing principles of an autonomous system may help align the system more closely with human values. This may also provide extra considerations for a prioritarian principle, rather than a maximisation principle, in evaluating trade-offs when normative conflicts arise.

A further difficulty is that normatively-relevant choices often require evaluation by comparing two competing interests. For example, it might be expected to prioritise harm prevention over human autonomy, or justice and fairness over predictive accuracy, or to place individual level interests above interests of the group or of society (Kearns & Roth, 2019). Not infrequently, a decision-maker is confronted with circumstances that create an obligation to do A and an obligation to do B, but in the circumstances they cannot perform both, that is, the decision-maker is faced by a so-called 'moral dilemma' (Brink, 1994). A second such account is where one valuable aim or thing cannot be achieved without causing damage or diminishment to another valuable one. For instances of such choice, we believe we need to say something about the features and factors - that is, the rationale and criteria informing the choice - that underlie the selection of placing one option above another in a given choice scenario. A selection that, we propose, is the product of, and has undergone, the moral scrutiny of the human decisionmaker/s (or stakeholder/s) and requires normative evaluation and deliberation involving ethically defensible moral choices, value judgements, compromises, and trade-offs.

We situate the process of using decision ladders to map normative choice within the idea of responsible technology, broadly construed, that is, that technology makes a contribution to, and promotes, human flourishing (Bynum, 2006). This interdisciplinary process demonstrates a practical way to refine and map human-and-agent interactions to facilitate normative conflict resolution that, ultimately, contributes to human happiness and well-being (Jirotka & Stahl, 2020). As such, this touches on how agents can make normative decisions responsibly, and with a degree of adaptivity, within the context into which they are deployed, that is, it forms part of a framework of responsible innovation that anticipates, reflects, engages, and responds (Stilgoe et al., 2020).

Although we do not engage directly with the difficult question of trust, increased multi-user stakeholder involvement and insight are supportive of trust in AI, which is positively associated with user and societal well-being (Choung et al., 2023). Trust in the system is a significant driver of adoption, and allows for better integration and acceptance in practice and for enhanced and personalised user experience (Townsend et al., 2023). Trust is also a cornerstone of ethical AI adoption as offered by various AI frameworks, such as the European Commission's High-Level Expert Group Guidelines for Trustworthy AI (European Commission, 2019). By developing this process, we further progress the issue of trustworthiness in autonomous agents as they are faced with normative decision-making in increasingly complex settings.

Decision-making has long been a focus of study within the field of psychology. Early theories and approaches to decision-making focused on option generation and selection, with suggestions that probability and utility estimates could be generated to decide between different options. Yet, the introduction of the Naturalistic Decision Making (NDM) field in the late 1980′s opposed this way of thinking about decisions and instead suggested ways in which people made decisions in the 'real world' (Klein, 2008). Prior laboratory studies had shown how people should make decisions, with tightly controlled parameters but they failed to capture how people actually make decisions in reality (Orasanu & Connolly, 1993). NDM research relies on field studies and subject matter experts to understand how people make decisions in more difficult conditions, with uncertainty and time restrictions. Through this it has been established that a key element within the decision-making process is experience, with novice and expert decision makers being classified on the level of experience they have with the environment and events that surround the decision (Klein, 2011; Canon-Bowers & Bell, 2014).

NDM aims to study the effectiveness of decision-making processes and provide guidance as to how people can make improved decisions, especially in difficult conditions. Numerous decision-making theories have been developed within the field of NDM, which share the inclusion of 'real world' elements of the decision making process, but vary in how they conceptualise decision-making (Lipshitz et al., 2001; Lintern, 2010; Parnell et al., 2022). There is no clear 'best' model to be applied, each naturalistic decision model will capture the individuals' perspective and the broader context surrounding the decision. A popular approach that has endured over time relates to the skills, rules, knowledge (SRK) levels of human performance that were proposed by Rasmussen (Rasmussen, 1983). The theory relates to the different levels of cognitive processing that are required to make sense of a situation and respond to it. Rasmussen (Rasmussen, 1983) states that different information is required at each of the SRK levels of performance and that these levels are hierarchical, such that behaviour at each of the levels increases in complexity.

Skill-based behaviour is automated and unconscious, it relates to the actions that are not intentional but have been well practised over time such that they become automated. There is a lack of higher level processing of behaviour at this level, and Rasmussen refers to this as "The man looks rather than sees" [(Rasmussen, 1983), p. 259]. Rule-based behaviour involves the 'stored rules', which are formed from previous experience of similar events or are given as instructions. Rules are guided by higher-level goals in an implicit manner, usually cued by the situation or environment and informed by previously successful behaviours. Skill- and rule-based behaviour are distinguished by the level of conscious attention involved, with skill based behaviour being unconscious and rule based behaviour being more intentional and informed by explicit know-how (Rasmussen, 1983). Knowledge-based behaviour is the highest level of performance wherein higher-level goals explicitly inform behaviour. A reliance on past experience is not enough and instead further information from the environment and situation must be processed to inform performance. This involves considering different options and reviewing possible outcomes by predicting different consequences. At this level, the mental model for the situation is consulted and informs future actions. Mental models contain an individuals' beliefs and understanding of the world, they are continually updated based on new and evolving events and information, and they guide an individuals' behaviour (Johnson-Laird, 1989).

The SRK levels of performance have been used to inform the design of human-machine interfaces (e.g. (Drivalou & Marmaras, 2009; Lin et al., 2011)) and levels of automation (Sheridan, 2017; Khastgir et al., 2018). Sheridan (Sheridan, 2017) states that automation performance has the same hierarchical SRK levels of performance as human performance. Standard feedback control relates to the skill-level, adaptive control relates to the rule-level, with rules governing behaviours in certain contexts. The knowledge-level relates to more advanced automation including Artificial Intelligence, deep learning and neural networks (Sheridan, 2017). The utility in this perspective is that the shared hierarchy between humans and automation can aid in the design of effective human-automation interaction. Performance can be attributed to the human and the automation across the different SRK levels. With regards to the focus of this paper, the SRK theory can determine which elements of normative choice can be implemented by the automated agent, and which should include input from human stakeholders. One means of doing this is through the application of the decision ladder method, which is a naturalistic decision model based on the SRK levels of performance, and is one method from a broader toolkit of methods encompassing Cognitive Work Analysis (Vicente, 1999; Jenkins et al., 2017).

3. The decision ladder

The decision ladder is a method within the CWA framework that aims to map out the system's decision points by providing the inputs, possible options available and the required processes to action the selected option. A diagram of a generic decision ladder is shown in Fig. 1.

The 'ladder' is composed of two streams that feed into and out of the top section. The diagram in Fig. 1 should be read from the bottom left 'Activation' element, directing up the ladder to the 'goal' before going down the right side of the ladder and finishing at the 'execute' element. The left side of the ladder is concerned with situation analysis which presents the conditions for the decision, including being first 'alerted' to the situation and processing the 'information' in the environment that determines the current state of the system. The decision is made at the top of the ladder, with the different 'options' considered in relation to the top-level 'goal' of the system. The selected option then sets in motion the 'tasks' and 'procedures' that are required to enact this option, feeding down the right-hand stream of the ladder, ending with the 'execute' point. The process of going up and down the ladder relates to the different levels of processing across the SRK-levels of performance, as annotated on the left of the ladder in Fig. 1. Skill-based behaviour is involved in the initial 'Activation' of the decision-making process, through the interpretation of cues or triggers from the environment. Skill-based behaviour is also implicated at the end of the decision process, at the bottom of the right-hand side of the ladder, when the behaviours that are directed by the decision are executed. Rule-based behaviour is also evident on both legs of the ladder. On the left-hand side, it is employed to make sense of environmental information, encoded through the prior skill-level processing, to diagnose that a decision will be required to ensure effective future behaviour. On the right-hand side of the ladder, rule-based behaviour is employed to determine the best course of action in order to enact the decision, including determining what tasks will be required and how they will be carried out. The top section of the ladder involves knowledge-based behaviour wherein options for possible future courses of actions are generated and compared to the top-level goal of performance to determine the best option.

The boxes within the ladder contain 'information processing activities' and the circles contain 'states of knowledge' resulting from the outputs of these activities. The metaphor of the ladder captures the incremental process of obtaining information from the environment, building information about the situation and deciding how to proceed, before going back down the ladder to enact the decision and adjust behaviour accordingly. However, the process of going up and down the ladder is informed by the level of experience of the decision maker. Experience can be built up through exposure to the environment or situation that a decision occurs within. Expert decision-making is represented through shortcuts across the ladder, without the need for more extensive deliberation at the knowledge base levels of behavioural control. This is because expert decision-makers hold good knowledge of

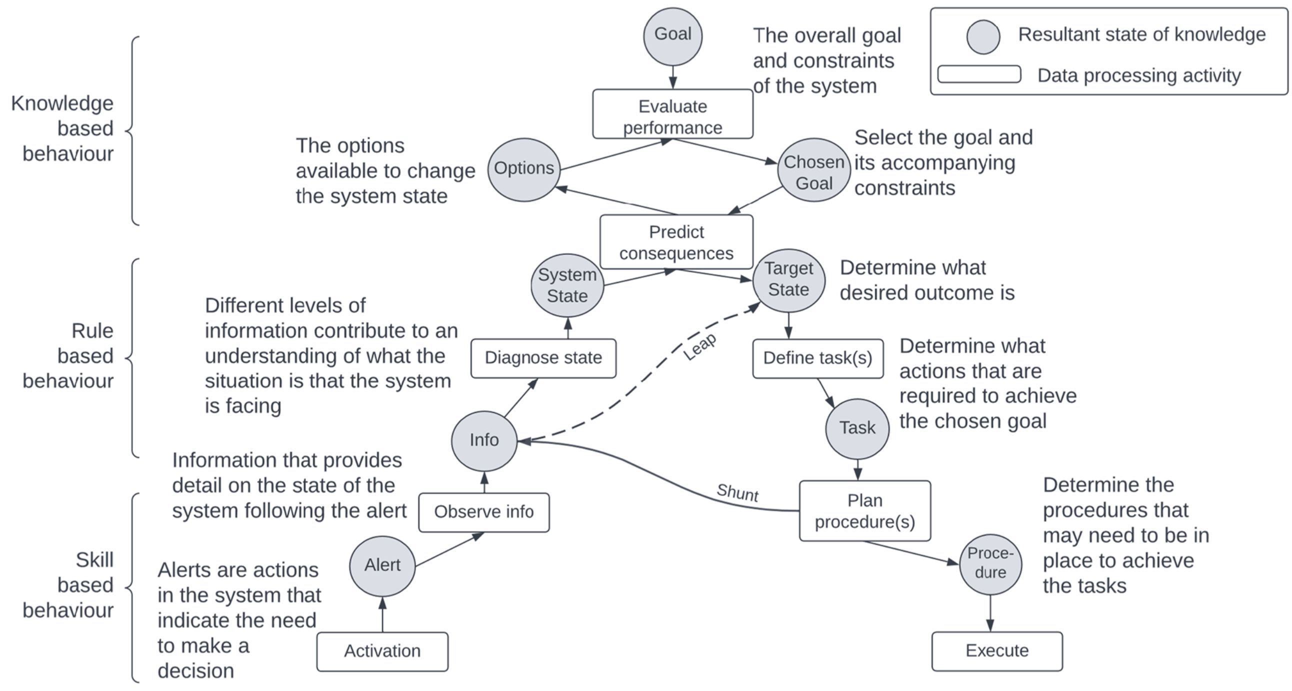
Knowledge based behaviour
Rule based behaviour
Skill based behaviour
Goal
The overall goal and constraints of the system
Resultant state of knowledge
Data processing activity
Evaluate performance
The options available to change the system state
Options
Chosen Goal
Select the goal and its accompanying constraints
Predict consequences
System State
Target State
Determine what desired outcome is
Different levels of information contribute to an understanding of what the situation is that the system is facing
Diagnose state
Leap
Define task(s)
Determine what actions that are required to achieve the chosen goal
Task
Info
Information that provides detail on the state of the system following the alert
Observe info
Shunt
Plan procedure(s)
Determine the procedures that may need to be in place to achieve the tasks
Alert
Proce-dure
Alerts are actions in the system that indicate the need to make a decision
Activation
Execute

Fig. 1. Overview of the decision ladder.

the system and can utilise this knowledge to inform their decision response without needing to review multiple options. This is shown through the arrows connecting the left and right streams of the ladder through leaps and shunts. Shunts refer to shortcuts between data processing to knowledge states wherein an expert can infer understanding about the system from the information that they are processing about it. Leaps refer to shortcuts between two knowledge states whereby the expert can infer information about the system based on their understanding of other aspects of the system.

Shortcuts from the left to the right stream of the ladder show where familiarity and expertise trigger action and shortcuts from the right to left steam show where desired actions require further information from the environment. Novice decision making requires more deliberation at the knowledge-based stages of the decision process, including determining what options are available and how they relate to the high-level goal of the systems (Rasmussen & Jensen, 1974; Vicente, 1999). This is also true of expert decision making in unfamiliar situations where they have less expertise and require more considered decision making processes (Jenkins et al., 2010).

The decision ladder offers an opportunity to map collaborative human-automation decision making through reviewing the behaviours across the SRK levels of control. As we are interested in the involvement of human stakeholders in the moral decision enacted by autonomous agents, the level of input from stakeholders can be determined through the level of processing that they contribute to the decision. Humans currently have better awareness of the morals and norms that shape effective behaviour in relation to any given situation compared to autonomous agents. Therefore, it is important to include human judgement within the knowledge-based levels of processing in these types of decisions. Yet, automation is now competent in performing skill-level behaviours through effective feedback control mechanisms and triggers that respond to specific cues within the environment (e.g. (Haidekker, 2020; Ward, 2000)). For example, when driving a vehicle, the human presses the brake pedal which will automatically initiate the feedback mechanism to engage the brake light on the vehicle. This is not to say that the skills of a human are the same or equivalent to those of the autonomous system, as humans and autonomous systems operate through different mechanisms. However, the outputs of the behaviours are comparable and when reviewing the increasing complexity of autonomous agents this is a useful comparison to make. Automation is also increasingly competent at performing adaptive feedback at the rule-level of behaviour (Sheridan, 2017). Taking the vehicle example again, adaptive cruise control technology is an advanced driver assistance technology that monitors vehicles on the road ahead and automatically adjusts the vehicles speed to keep a safe distance. Here, rule-based behaviour of maintaining a safe distance behind the vehicle is employed. Higher-level functioning of vehicle automation, whereby vehicles engage in knowledge-based processing is not yet available - largely due to the ethical, moral and legal issues (discussed in the previous section) that require end-user and stakeholder input (Keeling, 2020; Siegel & Pappas, 2021).

Our approach is to capture the factors and features that inform the choice and the decision-making in a decision ladder. These ‘features’ are the ethically relevant features in a particular case of normative decision-making that inform the appropriate course of action (Anderson & Anderson, 2015). When the underlying normatively-relevant feature(s) change(s), a change in action or outcome may be justified and a new course of action selected. This speaks to both the resilience and adaptability of the agent. Not only can the context and domain (for a time, at a place, and for a culture), and the needs and requirements change, but the underlying premises upon which a decision is made are not fixed and are themselves open to re-evaluation.

Once the factors and features have been captured, the task is to determine the overall normative status of these factors or features by establishing how they might combine and interact with one another (Kagan, 1988). Kagan (Kagan, 1988) argues that for the most part, the role of these factors ‘in determining the overall moral status of an act simply cannot be adequately captured in terms of separate and independent contributions that merely need to be added in’. The combination of two or more factors is not the sum of their independent contributions – although each of the positive factors provides a reason for performing the act and each of the negative factors provides a reason against selecting a choice. Choices based on simply tallying the pros and cons of the factors also do not tell us anything about what counts as a factor, about its strength or importance relative to another within a context, or about its reliability. The factors are contextually dependent, some have more importance than others - while others may have little or no value at all - and their value is not necessarily interchangeable between contexts.

We suggest that methods of documentable, accountable deliberation with stakeholders will assist to establish these contributing factors and help to resolve the trade-offs between the conflicting principles. Based on the results of the human decision-makers – expressed as a choice selection – we can then map out the decision-making process and create a generalised rule or precedent – one that systematically covers the domain of the possible, informed by typical examples and, from this, interpolate any new situations in terms of the existing known selected results. The selection informs the specificity of explicit concrete rule formulation.

4. Defeasible rules and defeaters

Stakeholders are invited to write normative rules together with ‘defeaters’ as informed by the decision ladder. Once defeasible rules are identified, non-monotonic logic is used to introduce any counter-examples that may challenge the validity of the rule, that is, in the form of a ‘defeater’. The defeater is the hedging-clause (or the ‘unless’ portion of the rule) and sets out the conditions and circumstances under which the default or defeasible normative rule does not hold. Thus, a normative rule has the general form ‘when A then B unless C’, for example, when a user refuses a non-critical medication, then do not administer the non-critical medication (that is, respect the user’s autonomy) unless refusal is repeated a maximum number of times whereupon report such repeated refusal to a human supervisor.[1]

As certain default rules have higher priority – or salience – than others, based on this salience these may be priority ordered and executed unless overridden by either an exception (in the form of a defeater) or a new normative rule. The presumption then is that a rule applies unless on the facts it is excluded by virtue of another rule with greater priority. Accordingly, a default rule is defeated in a decision-context when a stronger default which supports a conflicting conclusion is triggered in the scenario. In this way we can allow for the use of reasoning in a process in which the system can draw plausible and tentative, but not infallible, conclusions that can subsequently be retracted based on further evidence. This creates a mechanism of revising rules in the face of the acquisition of new information and which, in the right circumstances and within a context, stands to defeat another.

The difficulty with this, however, is that this framework does not provide assurance that every and all possible defeaters have been identified, only that a methodology is in place to accommodate further defeaters should the reasons for introducing them arise. It also still remains necessary to explain how (or whether) the system should rank all rules pari passu (on an equal footing) or in a particular default priority.

[1] We note that this normative rule can also be written ‘if (A and not C) then B’. However, based on recommendations we received from autonomous-system stakeholders including lawyers, ethicists, sociologists and psychologists (e.g., see [54, 55, 56, 57]), we opted for the format used in the paper. In addition to supporting stakeholder needs, this format improves the readability of rules with multiple defeaters.

In ranking, for instance, and as we will see in the example in Fig. 3 of the case study described in section 5 below, even the obvious reason - to not do harm - above, for instance, respecting human autonomy, we cannot be sure that the harm is not minor or negligible in a particular situation in comparison to a potential and grave infringement on human autonomy, for example. This is because the rule (and the underlying reasons for introducing it) and its salience is contextually sensitive and dependent, carrying practical relevance within a specific decision context.

As described above, these rules, defeaters, and priority relations are decided on the reliability and specificity of epistemic knowledge and human reasoning of the stakeholders. That is, we rely on an intuitive appraisal and the reflective judgement of certain practical, reasonable, and knowledgeable persons – or, 'humans-in-the-loop' – often domain experts, ethicists, members of the public, and stakeholders who are skilled in the art of making normative decisions within the particular use case. Accordingly, it is stakeholders that, using the aid of the decision ladders, provide the content of the rules and the defeaters, and their priority ranking.

Following this, such rules and defeaters are written in natural language and translated into specification, operational rules. To illustrate this, a case study of a carebot is described next that outlines conflicting values which require resolution using a combined autonomous agent and stakeholder process. The decision ladder is applied and developed to show how such conflicts can be resolved through mapping the stakeholders and autonomous agents capabilities and involvement.

5. Case study of a carebot

A carebot is required to select a course of action either supported by the principle of respect for autonomy or one underpinned by nonmaleficence (preventing harm). Suppose that a user exercises their right to human autonomy (by refusing an action be performed, such as

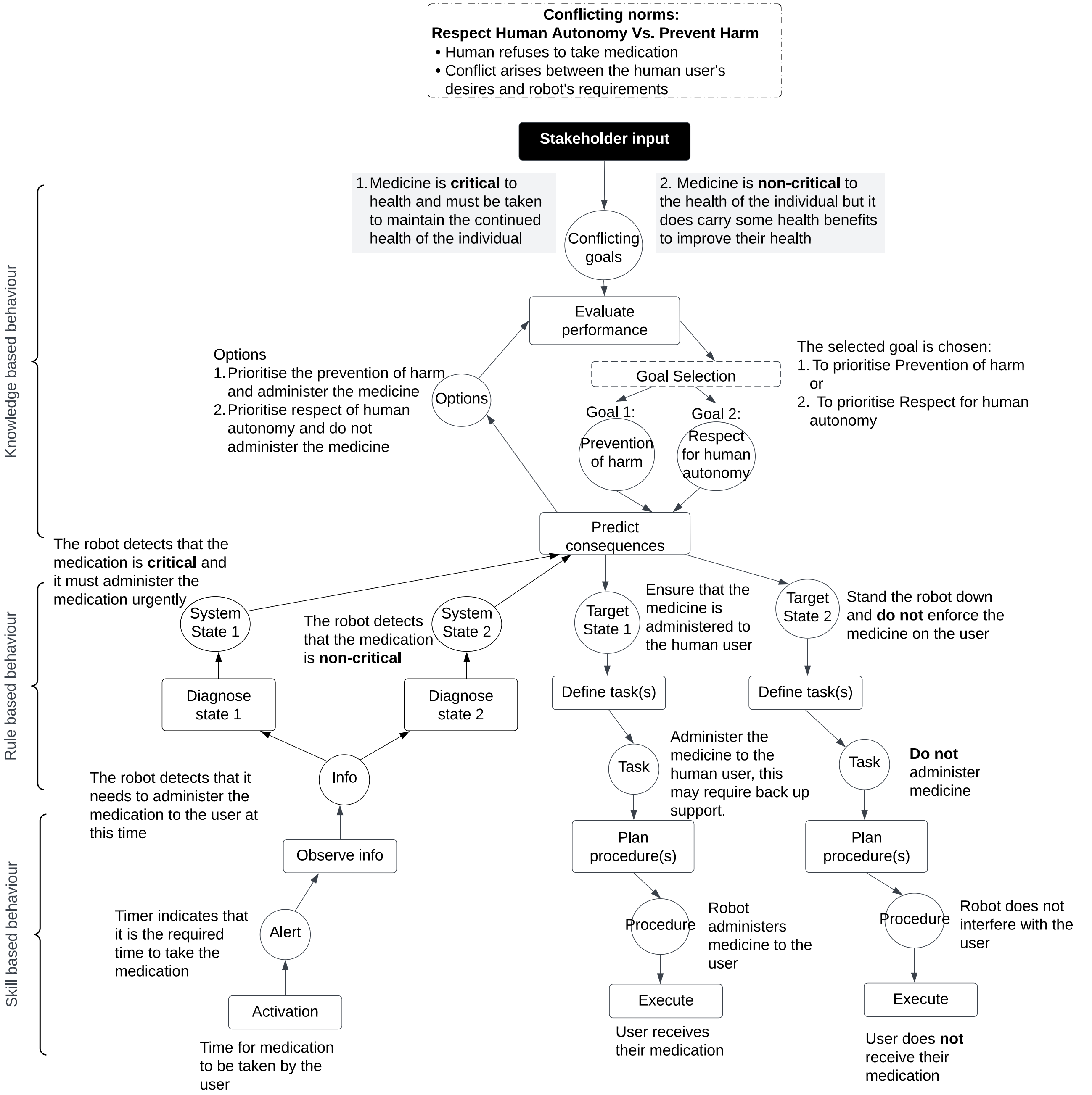
Conflicting norms:
Respect Human Autonomy Vs. Prevent Harm
• Human refuses to take medication
• Conflict arises between the human user's desires and robot's requirements
Stakeholder input
1. Medicine is critical to health and must be taken to maintain the continued health of the individual
2. Medicine is non-critical to the health of the individual but it does carry some health benefits to improve their health
Conflicting goals
Evaluate performance
Options
1. Prioritise the prevention of harm and administer the medicine
2. Prioritise respect of human autonomy and do not administer the medicine
Options
Goal Selection
The selected goal is chosen:
1. To prioritise Prevention of harm
or
2. To prioritise Respect for human autonomy
Goal 1:
Prevention of harm
Goal 2:
Respect for human autonomy
Knowledge based behaviour
Predict consequences
The robot detects that the medication is critical and it must administer the medication urgently
System State 1
The robot detects that the medication is non-critical
System State 2
Target State 1
Ensure that the medicine is administered to the human user
Target State 2
Stand the robot down and do not enforce the medicine on the user
Rule based behaviour
Diagnose state 1
Diagnose state 2
Define task(s)
Define task(s)
The robot detects that it needs to administer the medication to the user at this time
Info
Task
Administer the medicine to the human user, this may require back up support.
Task
Do not administer medicine
Observe info
Plan procedure(s)
Plan procedure(s)
Skill based behaviour
Timer indicates that it is the required time to take the medication
Alert
Procedure
Robot administers medicine to the user
Procedure
Robot does not interfere with the user
Activation
Execute
Execute
Time for medication to be taken by the user
User receives their medication
User does not receive their medication

Fig. 2. Decision ladder that presents the conflicting goals. The two downward options on the right show the different options for resolving the conflict.

getting dressed or taking their medicine) but in doing so exposes themselves to harm. Respecting human autonomy and preventing harm inform the execution of divergent courses of action, and both have identifiable courses of action that are ethically indicated in the circumstances.

Trading-off two 'competing' interests (autonomy and non-maleficence) within a dynamic context will depend on, amongst other things, the importance of the respective interests, the relative importance of each alternative, the importance or salience/significance of any losses and gains, the degree of severity of consequence for non-selection, and an assessment of the certainty of the underlying assumptions made and/or of the likelihood of a given outcome. A Cognitive Work Analysis and a decision ladder offer the opportunity to review these elements while considering the broader context within which the decision is made.

The decision ladder method is applied to the carebot scenario to show its value in providing insights into the decision-making process, including the human-robot interactions and the wider environmental context. As we are considering the two competing norms of 'respecting human autonomy' and 'non-maleficence/prevention of harm' within this scenario, the criticality of the medication to the health of the individual is a key decision factor that must be included within the decision to administer the medication to the user who refuses it or not. Medication such as insulin, for example, is health critical to an individual with diabetes and it must be taken routinely otherwise the user's health may significantly suffer. Other medications are less health critical, for example taking a vitamin tablet or health supplement. While these may help to support the health of the user, their health will not significantly deteriorate in the short term if these medications are not taken. Within this scenario we refer to the user as the person who requires to take the medication.

The decision ladder in Fig. 2 provides an overview of the decision process in administering the medication to the user with respect to the environmental factors of the medication criticality and the norms of personal autonomy and non-maleficence. To capture the impact of conflicting norms on the decision making process, the decision ladder method is expanded upon with additional downward ladders representing the different outputs of the conflicting principles.

The initial starting point is the carebot notifying the user that they need to take their medication, which starts at the skill-based level of behaviour and moves up to the rule-based level of behaviour. The 'Activation' element on the bottom left of Fig. 2 feeds up to the 'Alert' which is the state of knowledge event (denoted by the circle outline) that arises from the timer notifying the user that it is time for the medication. This happens at the skill level, a simple trigger of the timer which can be readily automated and interpreted. The 'Observe info' event is an information processing activity (denoted by the box outline) which shows that the alert needs to be processed to determine what it is trying to communicate to the system. An automated system is capable of interpreting such information, processing the alert and its triggered action. The carebot can understand that the alert means that the user must take their medication and notifies that user of this, as shown by the 'Info' knowledge event in Fig. 2.

At this point the agent will likely need to interact with the user to inform them that they need to take their medication. The mechanisms for this interaction are beyond the scope of this paper, however there is progressive research in human-robot interaction (Weiss & Spiel, 2022). Importantly this interaction will need to be context aware (Liu & Wang, 2021; Quintas et al., 2018) and large language models may be able to help analyse concepts similarity across multiple normative requirements (M. L. Y. S. S. I. B. Y. A. R. d. M. V. T. B. B. H. C. A. Feng, N., R. Calinescu, 2024). This research does not focus on this detail but is more concerned with the high-level normative conflict which may arise from such interactions. The processing of this information by the agent and the user may initiate the basis for conflict to arise and the user may refuse the medication, choosing to ignore or prevent the carebot from administering the medication. At this point the system identified that there are two possible states that can diagnosed when the patient is refusing their medication. In case 1 the patient is refusing critical medication which will have negative consequences for their health. In the second case the medication is not critical but is still beneficial to their sustained overall health. The 'Diagnose state' is the information processing activity that involves that agent and the user interacting and the agent determining that the user may not be willing to take their medication. The conflict arises between the agent needing to administer the medication to 'prevent harm' (non-maleficence) from the user not taking their medication, and to allow the user to maintain 'human autonomy' to make their own choices and refuse the medicine. Fig. 2 shows that this results in the conflicting norms being prevalent at the 'system state' level, which is the highest level of the rule-based behaviours on the decision ladder. Fig. 2 shows two system states, depicting case 1 where the medicine is critical and state 2 where it is not critical. The carebot should be able to obtain an assessment of the system state by comparing what is meant to occur in the event of the 'alert' and how that varies from the current behaviour of the system. To resolve this divergence, and understand the role that normative values have on the decision-making process, knowledge-based behaviour is required.

The knowledge-based behaviour section at the top of the decision ladder is presented as a loop, beginning and ending with the information processing activity 'predict consequences'. This refers to the idea that decision makers review the consequences of their action within their decision-making process and use this to guide them. This concept is prevalent in many popular naturalistic decision making theories, for example the recognition primed decision model (Klein, 1993) involves the mental simulation of possible actions to review their effectiveness. In the decision ladder, the current system state is reviewed with higher-level knowledge based behavioural processing to determine what the consequences are of the current course of action. It then uses this to review possible 'options' in relation to the top-level 'goal' of the system. The 'options' shown in Fig. 2 show the options of prioritising each of the conflicting norms. At the knowledge level, the broader complexities of the norms are realised and considered relative to the performance of the system itself. Such behaviour should be delegated to human decision-makers as 'responsible/reasonable persons' as stated earlier in the paper. Within the decision ladder the options are evaluated with respect to the goals of the system. The norms that are in conflict with each other relate to different and conflicting higher-level systems goals. To resolve the conflict, input from stakeholders is required to understand how these goals should be prioritised. This is shown in Fig. 2 as the top-down process influencing the conflicting goals at the very top of the ladder. Engagement with stakeholders such as health practitioners, carers and patients can provide insight into how such priorities should be resolved. This is not to say, however, that stakeholder input cannot inform other elements of the decision ladder.

The top-down input from the knowledge based processes by stakeholders will feed down through the descending leg of the ladder to inform the best course of action in response to each target state. Yet, stakeholder input is also needed on the ascending leg of the decision ladder to inform how the robot should interact with the human user to inform them that they need to take their medication, as well as how the robot should respond when the patient refuses to take their medication. Stakeholder engagement in such interactions are covered elsewhere in the literature e.g. (Weiss & Spiel, 2022).

Within this scenario, it is highly important that the context of the situation is included within the decision making process. The time dependency of the medication is a key contextual factor that will influence the decision and outcome of the scenario. Fig. 2 shows the importance of the criticality of the medication to the decision making process, feeding down from the top level goal of the decision process. The time criticality of the situation will influence which conflicting goal will be selected and whether to prioritise human autonomy or to prioritise the prevention of harm.

Ordinarily, the decision ladder has one down-ward leg to capture the output of the decision and the required action needed to enact the chosen option. The decision ladder in Fig. 2 has two downward legs which capture both possible outputs from the decision, which are selected depending on the criticality of the medication. These are informed by the two diagnosed states from the ascending leg of the decision ladder. Goal 1 relates to the prioritising of the prevention of harm, in the case that the medication is critical and the user takes it as soon as possible to maintain their continued health. Goal 2 involves prioritising human autonomy, wherein the medication is not critical and can be re-reviewed in some hours or followed up later. The two processes for carrying out either goal are shown on the decision ladder through the two downward legs. There are alternative options, and only one will occur at a time. Once the option is selected and the rule-based behaviour takes over once again to determine what the target state of the system is. This target state is a contrast to the current system state that was previously determined on the left-side of the ladder. Determining the target state will lead to a set of tasks and procedures to be defined to establish this state. Following ‘Target state 1′, the task is to administer the medication to the user, which may include calling for backup support. For ‘Target state 2′, the carebot stands down and does not administer the medication. The high-level tasks invoke finer-detail procedures which are considered at the skill-based level of behaviour.

The decision ladder clearly maps out where stakeholder involvement is required within the knowledge-based behaviour level of the decision-making process. Their input can help to understand where certain values need to be prioritised as well as where further support may be needed to assist with the decision-making process. The decision ladder also shows how stakeholder input and robot functionality can work in collaboration with each other following the SRK levels of behaviour.

Fig. 2 shows the process for enacting a decision once the goal and target state have been identified and no further challenges arise. However, as suggested in (Townsend et al., 2022), norms and operationalised rules can be refined and extended by generating possible defeaters. Once rules have been identified, the conditions wherein a rule can be ‘defeated’ can be generated to show how conflicts may be resolved at the rule-based level. In other words, by understanding where possible rules may need to be reconsidered in favour of other alternatives, the strength of the system as a whole can be considered. Furthermore, the defeaters to the rules can consider additional normative concerns which may impact on the effectiveness of the autonomous agents’ behaviour.

Fig. 3 presents a decision ladder that builds on that presented in Fig. 2, by including defeaters. This figure is an extend version of Fig. 2, providing more in-depth detail into the decision making process with respect to defeaters. The defeater suggests that more information relating to the timing of the medication is also important to the decision. In Fig. 3 two different information states are provided to represent the information relating to the criticality of the medicine to the patients health as well as the time criticality of the medication. This leads to four possible diagnosed states. Information branch 1 states that the medicine is health critical and therefore the patient will need to take the medication. The different diagnosis states provide different time critical diagnoses. Diagnosis state 1a states that the medication much be given immediately to maintain the patients health. Yet, diagnosis state 1b states that while the medication is health critical it is not immediately needed but must be administered within the next two hours. This time limit suggests that the medication can be attempted to be re-administered at any time within the proceeding two-hour window. Information state 2 states that the medicine is not health critical. There are two further diagnosis states here, Diagnosis state 2a implies that the

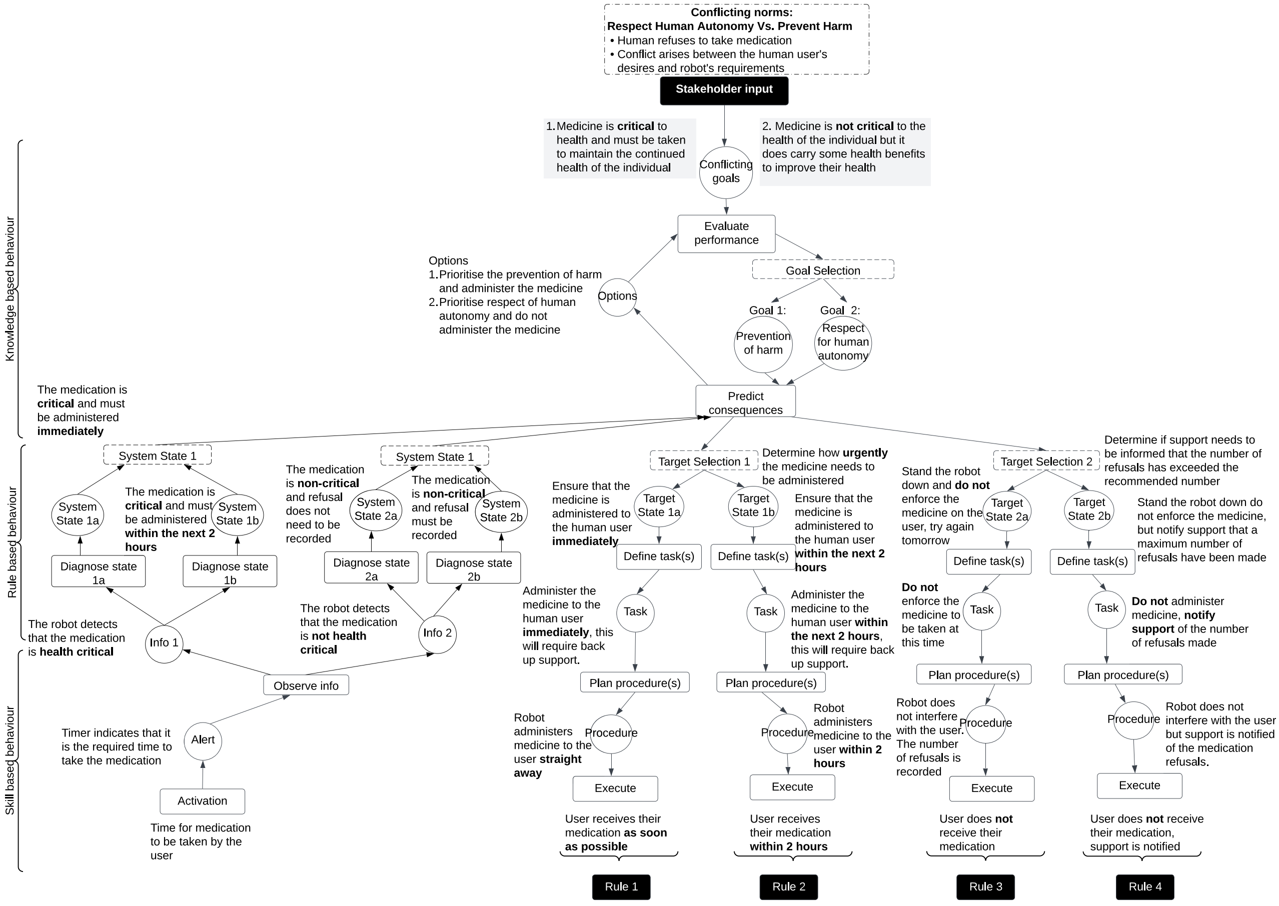
Conflicting norms:
Respect Human Autonomy Vs. Prevent Harm
• Human refuses to take medication
• Conflict arises between the human user's desires and robot's requirements
Stakeholder input
1. Medicine is critical to health and must be taken to maintain the continued health of the individual
2. Medicine is not critical to the health of the individual but it does carry some health benefits to improve their health
Conflicting goals
Evaluate performance
Goal Selection
Goal 1:
Prevention of harm
Goal 2:
Respect for human autonomy
Options
1. Prioritise the prevention of harm and administer the medicine
2. Prioritise respect of human autonomy and do not administer the medicine
Options
Predict consequences
Knowledge based behaviour
Rule based behaviour
Skill based behaviour
The medication is critical and must be administered immediately
System State 1
System State 1a
The medication is critical and must be administered within the next 2 hours
System State 1b
Diagnose state 1a
Diagnose state 1b
The robot detects that the medication is health critical
Info 1
System State 1
The medication is non-critical and refusal does not need to be recorded
System State 2a
The medication is non-critical and refusal must be recorded
System State 2b
Diagnose state 2a
Diagnose state 2b
The robot detects that the medication is not health critical
Info 2
Observe info
Timer indicates that it is the required time to take the medication
Alert
Activation
Time for medication to be taken by the user
Target Selection 1
Determine how urgently the medicine needs to be administered
Ensure that the medicine is administered to the human user immediately
Target State 1a
Target State 1b
Ensure that the medicine is administered to the human user within the next 2 hours
Define task(s)
Define task(s)
Administer the medicine to the human user immediately, this will require back up support.
Task
Task
Administer the medicine to the human user within the next 2 hours, this will require back up support.
Plan procedure(s)
Plan procedure(s)
Robot administers medicine to the user straight away
Procedure
Procedure
Robot administers medicine to the user within 2 hours
Execute
Execute
User receives their medication as soon as possible
User receives their medication within 2 hours
Rule 1
Rule 2
Target Selection 2
Determine if support needs to be informed that the number of refusals has exceeded the recommended number
Stand the robot down and do not enforce the medicine on the user, try again tomorrow
Target State 2a
Target State 2b
Stand the robot down do not enforce the medicine, but notify support that a maximum number of refusals have been made
Define task(s)
Define task(s)
Do not enforce the medicine to be taken at this time
Task
Task
Do not administer medicine, notify support of the number of refusals made
Plan procedure(s)
Plan procedure(s)
Robot does not interfere with the user. The number of refusals is recorded
Procedure
Procedure
Robot does not interfere with the user but support is notified of the medication refusals.
Execute
Execute
User does not receive their medication
User does not receive their medication, support is notified
Rule 3
Rule 4

Fig. 3. Decision ladder with the defeaters for each of the conflicting goals, resulting in four possible rules.

medicine is not health critical and if it is not taken no further action is needed. Yet Diagnosis state 2b implies that while the medication is not health critical it is important to note down each refusals to determine how many times the patient has not taken the medicine. Similar conflict goal activity occurs at the knowledge based behaviour elements of the decision ladder as is Fig. 2, with the normative values of ’Respecting human autonomy’ and ’Preventing harm’ conflicting with each other. Each of the target states now has a defeater and therefore there are four downward legs of the ladder. Each of the downward legs presents an alternative course of action and again only one will occur at any one time. The different target states relate to the systems states on the ascending leg of the ladder. The different events that are triggered by the target states can then be translated into natural language rules, and then into a domain-specific language for the formal specification of social, legal, ethical, empathetic and cultural (SLEEC) rules (S. Getir Yaman et al., 2023), which will be discussed in the following section.

It should also be noted that the activation and alert at the skill levels are triggered in the same way and the different states are not realised at this base level of the ladder. Likewise, the two goals at the knowledge-based levels of the decision ladder are the same in Fig. 3 as they are in Fig. 2, involving a conflict between the goal of preventing harm and the goal of respecting autonomy. These goals are to be prioritised based on the criticality of the medication and it is through stakeholders input that the priority will be informed. ‘Target state 1a’ and ‘Target State 1b’ follow the route of action when prevention of harm is prioritised in the event of health-critical medication. In ‘Target State 1a’, the medication must be taken immediately and the tasks and procedures for administering the medication are then followed, calling for backup support where this may be required. However, a defeater is introduced for ’unless the medication is not immediately time critical’, for example if the medication is needed within the next 2 h. Therefore the tasks that follow in ’Target state 1b’ are less time critical but the carebot must seek to administer the medication or call for backup support within the given time frame (i.e. two hours).

‘Target State 2a’ and ‘Target State 2b’ follow the target state when the medication is not critical and therefore human autonomy is prioritised. ‘Target State 2a’ follows a similar process to ‘Target State 2′ in Fig. 2, the robot does not administer the medicine and they allow the human user to continue with their activities. A defeater has been included here to determine when support may need to be called to inform medical professionals that the medication has not been taken. Hence, in ‘Target State 2b’ the defeater is introduced as ‘unless the user refused medication several days in a row’, to identify the need to count how many times the user has refused the medication and to contact backup support if the number of attempts has reached the maximum.

Presenting the normative conflicts on the decision ladder in this way shows the collaborative efforts required by stakeholders at the knowledge-based levels, and by the carebot at the rule- and skill-based levels. Furthermore, it provides an overview of the tasks and procedures that the carebot needs to complete for each eventuality. Furthermore, presenting them in this way allows for the processes to be mapped and specified in the SLEEC language, as detailed in the next section.

It should be noted that there are a number of challenging environmental considerations that could significantly influence the outcomes and effectiveness of the carebot scenario. Environmental monitoring and response is a key area for robotic and autonomous system development and therefore the processes presented here should be interpreted with caution. However, they do give an overview of the key interaction activities between the carebot and the human patient from which additional environmental considerations can be made. Through engaging with stakeholders using this approach, the environmental considerations that experts in the area deem to be important can be captured.

6. Specification of rules

This section explains how rules regarding normative conflicts arising from the decision ladders can be written. The target sections in the decision ladder in Fig. 3 advise how to construct these rules by following the Target state from its Task to Procedure. Specifically, ‘Target State 1a’ implies a default rule in case medication is critical (e.g. medication is insulin or an inhaler) in the form

“Rule1: When the medication is ‘critical’ and the urgency is ‘high’, then administer the medication and call support immediately.”

To encode such a rule in a way that can be processed by an autonomous agent, we use a domain-specific language called SLEEC language, which is a rule-based, timed language for the formal specification of normative requirements introduced by Get,ir Yaman et, al. (S. Getir Yaman et al., 2023; S. Getir Yaman et al., 2023). A SLEEC rule defined in this language includes a trigger event which specifies the circumstances under which the rule applies. In this case, the rule is triggered by user input leading to a ConflictIdentified event. The other circumstances that the rule needs to take into account are information about whether the medication is critical, and whether the urgency or its administration is high. These types of information that the robot needs to be able to access are called measures in the SLEEC language. Each measure can have Boolean (i.e., True/False), numeric or scale type. For instance, the measure medicationIsCritical required to establish whether the robot should administer the medicine is a Boolean measure.

Given the trigger event and measures introduced so far, our rule defines the response required when a conflict is identified and the medication is critical: this response specified by means of a (required) event AdministerMedicationAndCallSupport - is that the robot should administer the medication and call support immediately.

‘Target State 1b’ implies a related rule which considers the case where the medication is critical but the urgency (specified as a scale measure that can have one of the val-ues high, medium and low) is not high (e.g., medication is an antibiotic instead of an inhaler for asthma patients, which requires immediate administration). A rule can be extracted from this implication in the form

“Rule2: When the medication is ‘critical’ and the urgency is ‘moderate’ or ‘low’, then stand off for 2 h before administering medication and calling support”.

Rule2 in this case, handles the exceptional condition as a defeater to Rule1. Hence, a combined rule of Rule1 as a default and Rule2 as a defeater can be encoded in the SLEEC rule-based language (using the unless language construct), as shown in Listing 1 where the combined Rule1and2 is defined under the assumption that “immediate time” is represented by 1 min.

A similar process can be applied to the ‘Target Selection 2′ from the decision ladder in Fig. 3. This section of the decision ladder captures the situation where the medication is not critical (e.g., because it is a vitamin or other supplement). In this situation, ‘Target State 2a’ advises not to enforce the medication for the current time and try to offer the medication at another time (i.e. the following day). A SLEEC rule can be formulated from this as

“Rule 3: When medication is ‘non-critical’, do not administer medication, count the number of (consecutive) requests where medication was not administered.”

A defeater arises from ‘Target State 2b’ when the number of consecutive refusals to take this medication reaches a limit, e.g. 10, in which case the medication needs to be given by a support team. This scenario yields the rule

“Rule4: When medication is ‘non-critical’ and numberof-attempts is 10, then advise/call support, and reset the number-of-attempts counter.”

As a consequence, a combined rule that brings together Rule3 as a default and Rule4 as a defeater can be encoded in the SLEEC rule-based language as shown in Listing 2.

7. Discussion and future work

Building on previous work that demonstrated the importance of identifying high-level normative principles that guide human-robot

Listing 1

Rule extracted from the Decision Ladder in Fig. 3 for the goal 'Prevention of Harm'.

```
Event ConflictIdentified
  Event AdministerMedicationAndCallSupport
  Measure medicationIsCritical: boolean Measure urgency: (low, medium, high)
  Rule1and2 when ConflictIdentified and medicationIsCritical then
    AdministerMedicationAndCallSupport within 1 min unless urgency<medium then
      AdministerMedicationAndCallSupport within 2 h
```

Listing 2

Rule extracted from the Decision Ladder in Fig. 3 for the goal: 'Respect for human autonomy'.

```
Event ConflictIdentified
  Event AdministerMedicationAndCallSupport
  Event IncrementNumberOfRefusals
  Event CallSupportAndResetCounter Measure medicationIsCritical: boolean
  Measure urgency: (low, medium, high) Measure numberOfAttempts: numeric
  Rule3and4 when ConflictIdentified and not medicationIsCritical then
      IncrementNumberOfRefusals unless numberOfAttempts> = 10 then
      CallSupportAndResetCounter
```

interactions to inform lower-level, explicitly formulated SLEEC rules (Townsend et al., 2022; Parnell et al., 2023), our paper has considered conflicting normative principles and their resolution and operationalisation. To that end, we have introduced a structured approach to making normative choices, and to reviewing and resolving conflicting norms with stakeholder input. This allows for better understanding how and when a possible conflict and decision arises and establishing what might be done about it. Decision ladders have been proposed as a useful tool to map out the conflicting norms, and to provide a structured approach to capturing and reviewing how the conflicts can be resolved with respect to stakeholder input and autonomous system capabilities. This paper is based on a somewhat straightforward example in which the conflict arose due to the criticality of the medication for illustration purposes. However, in many applications, such differences could be more subtle and/or complex. For instance, the scenario encountered in practice might require assessment of the emotional or cognitive state of the user. Previous work has focused on an assistive robot to care for the elderly, children or those with disabilities of a cognitive or physical nature, providing dressing support such as assisting with putting on shoes or items of clothing (Jevtić et al., 2018), (Co¸sar et al., 2020) (Townsend et al., 2022). In such a scenario, conflicts may arise between different norms, such as privacy and well-being. Yet a similar process can be mapped out to show where the conflict may be triggered in the ascending ladder, e.g. the selection of specific items of clothing, and how different options for managing the conflict may be applied in the descending ladder, e.g. fall-back mechanisms, additional support or alternative clothing items. Using video imagery, posture recognition and speech recognition such robots could monitor the emotional state of the individual and dictate possible available options, as well as how future interactions of the robot should be conducted. The importance of involving stakeholders within the design and development of autonomous agents is highlighted when considering conflicting normative principles. Future autonomous agents must be designed with input from stakeholders and this requires methods that identify where stakeholder input can be added and what value stakeholders can add. Decision ladders provide a structure for identifying where stakeholders can add value through their experience of the domain and context within which future decisions will need to be made. The SRK levels of performance defined by Rasmussen (Rasmussen, 1983) define the hierarchy of information processing capabilities, from basic skill level performance up to advanced knowledge-based processing. While originally developed to capture human capabilities, the SRK levels can also be applied to autonomous agents and the hierarchy of their information processing capabilities (Sheridan, 2017; Khastgir et al., 2018). The decision ladder method is based on the SRK levels of processing, and therefore maps the decision process onto the skills, rules and knowledge based information that is used to make decisions. Autonomous agents are currently capable of skill-based and rule-based behaviour and are able to take on these elements of decision making, such as acknowledging that there is a need for a decision (skill level) and identifying the information that is relevant to the decision (rule based). However, they currently cannot undertake knowledge-based processing which requires a more advanced, knowledge-based level of processing. Knowledge-based processing within decision making requires an understanding of the wider context within which the decision is to occur, the alternative options and an ability to make prediction about future actions. For applications such as the carebot from our example, this type of knowledge-based processing is currently only found in experienced stakeholders such as trained medical professionals and carers. Through breaking the decision process down into the SRK levels of processing, the combined input of autonomous agents and stakeholders can be observed.

Future steps to advance our approach should involve obtaining stakeholder input to feed into these decision ladders. As the decision ladder has been used extensively within critical domains in the past, there are numerous examples in the literature of collecting stakeholder input to feed into decision ladders (e.g. (Asmayawati & Nixon, 2020; Jenkins et al., 2010). This has involved semi-structured interviews with domain experts to understand their decision-making process. Future work should follow these semi-structured interview methods, with additional onus on how they feel automated systems could best assist within the decision-making process.

The structure of the decision ladder also offers the opportunity for mapping the components of the decision process to natural language rules that can then be incorporated into the SLEEC domain-specific language. This mapping provides a clear process for auditing how the decision emerge and the information that is used to inform the decision, as well as the processes involved in carrying out the actions arising from the option chosen. As demonstrated within our paper, this structured approach can show how conflicts may arise and how stakeholder expertise can be combined into the human-carebot interaction.

In the carebot scenario, the stakeholders would be medical professionals such as doctors, nurses, and carers who are responsible for administering medication to patients and monitoring their health. These stakeholders undergo significant training to develop knowledge and experience in supporting patients and making informed decisions about their health. Such domain experts would therefore be able to make quick and effective decisions in the scenarios presented; however, the health sector is currently under a significant amount of pressure from increasing demand. Autonomous agents offer the opportunity to increase resources, alleviating this pressure (Townsend et al., 2023). Yet, within such a highly critical domain, the effective integration of autonomous agents is crucial and trust plays a significant role (Fischer et al., 2018; Townsend et al., 2023). Trust within the human-robot interaction is not captured within this model, and requires further research to understand how patients would trust receiving the decision output from the carebot and if clearer transparency about the involvement of stakeholders may increase the trust in the carebot.

8. Conclusion

We have demonstrated how decision ladders can be used to provide a structured approach to aiding normative choice in autonomous agents. Using defeasibility, we have shown how specification rules with defeaters can be written to operationalise such choice. The decision ladder approach shows that stakeholder input remains an important input in the development of autonomous agents, with knowledge based information required at the higher levels of information processing. Future work should build on this approach to develop an accompanying process for stakeholder engagement.

CRediT authorship contribution statement

Beverley Townsend: Conceptualization, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Katie J. Parnell:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Sinem Getir Yaman:** Conceptualization, Investigation, Methodology, Writing – original draft. **Gabriel Nemirovsky:** Conceptualization. **Radu Calinescu:** Conceptualization, Investigation, Supervision, Writing – review & editing.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work described in this paper was funded by the EPSRC project EP/V026747/1 'UKRI Trustworthy Autonomous Systems Node in Resilience'.

References

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15. –15.

Anderson, M., & Anderson, S. L. (2011). *Machine Ethics*. Cambridge University Press.

Anderson, M., & Anderson, S. L. (2015). Toward ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm. *Industrial Robot: An International Journal, 42*(4), 324–331.

Anderson, M., & Anderson, S. L. (2018). Geneth: A general ethical dilemma analyzer, Paladyn. *Journal of Behavioral Robotics, 9*(1), 337–357.

Anderson, M., Anderson, S. L., & Armen, C. (2004). Towards machine ethics. In *AAAI-04 Workshop on Agent Organizations: Theory and Practice* (pp. 53–59). San Jose, CA.

Anderson, M., Anderson, S. L., & Armen, C. (2006). MedEthEx: A prototype medical ethics advisor. *Proceedings of the National Conference on Artificial Intelligence, 21*(2), 1759.

Anderson, M., Anderson, S. L., & Berenz, V. (2018). A value-driven eldercare robot: Virtual and physical instantiations of a case supported principle-based behavior paradigm. *Proceedings of the IEEE, 107*(3), 526–540.

Asmayawati, S., & Nixon, J. (2020). Modelling and supporting flight crew decision-making during aircraft engine malfunctions: Developing design recommendations from cognitive work analysis. *Applied Ergonomics, 82*, Article 102953.

Brink, D. O. (1994). Moral conflict and its structure. *The Philosophical Review, 103*(2), 215–247.

Bynum, T. W. (2006). Flourishing ethics. *Ethics and Information Technology, 8*, 157–173.

Canon-Bowers, J. A., & Bell, H. H. (2014). *Training Decision Makers for Complex Environments: Implications of the Naturalistic Decision Making Perspective* (pp. 99–110). Psychology Press. Ch. 10.

Cervantes, J.-A., Rodríguez, L.-F., Lopéz, S., Ramos, F., & Robles, F. (2016). Autonomous agents and ethical decision-making. *Cognitive Computation, 8*, 278–296.

Choung, H., David, P., & Ross, A. (2023). Trust and ethics in ai. *AI & Society, 38*(2), 733–745.

Co¸sar, S., Fernandez-Carmona, M., Agrigoroaie, R., Pages, J., Ferland, F., Zhao, F., Yue, S., Bellotto, N., & Tapus, A. (2020). Enrichme: Perception and interaction of an assistive robot for the elderly at home. *International Journal of Social Robotics, 12*, 779–805.

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

Dodig Crnkovic, G., & C¸ürüklü, B. (2012). Robots: Ethical by design. *Ethics and Information Technology, 14*, 61–71.

Drivalou, S., & Marmaras, N. (2009). Supporting skill-, rule-, and knowledge-based behaviour through an ecological interface: An industry-scale application. *International Journal of Industrial Ergonomics, 39*(6), 947–965.

Driver, J. (2005). Normative ethics. *The Oxford Handbook of Contemporary Philosophy*, 31–62.

European Commission. High-level expert group on artificial intelligence, ethics guidelines for trustworthy AI (2019), URL {https://digital-strategy.ec.europa.eu/en/library/.ethics-guidelines-trustworthy-ai}.

Fischer, K., Weigelin, H. M., & Bodenhagen, L. (2018). Increasing trust in human–robot medical interactions: Effects of transparency and adaptability, Paladyn. *Journal of Behavioral Robotics, 9*(1), 95–109.

Getir Yaman, S., Burholt, C., Jones, M., Calinescu, R., & Cavalcanti, A. (2023a). Specification and validation of normative rules for autonomous agents. In *26th International Conference on Fundamental Approaches to Software Engineering* (pp. 241–248). Springer-Verlag. https://doi.org/10.1007/978-3-031-30826-0 13.

S. Getir Yaman, A. Cavalcanti, R. Calinescu, C. Paterson, P. Ribeiro, B. Townsend, Specification, validation and verification of social, legal, ethical, empathetic and cultural requirements for autonomous agents (July 2023b). arXiv:2307.03697. URL https://arxiv.org/abs/2307.03697.

Getir Yaman, S., Ribeiro, P., Burholt, C., Jones, M., Cavalcanti, A., & Calinescu, R. (2024). Toolkit for specification, validation and verification of social, legal, ethical, empathetic and cultural requirements for autonomous agents. *Science of Computer Programming, 236*, Article 103118.

Haidekker, M. A. (2020). *Linear Feedback Controls: The Essentials*. Elsevier.

Horty, J. (2012). *Reasons as Defaults*. Oxford University Press.

Jenkins, D. P., Stanton, N. A., Salmon, P. M., Walker, G. H., & Rafferty, L. (2010). Using the decision-ladder to add a formative element to naturalistic decision-making research. *International Journal of Human–Computer Interaction, 26*(2–3), 132–146.

Jenkins, D. P., Stanton, N. A., & Walker, G. H. (2017). *Cognitive Work Analysis: Coping With Complexity*. CRC Press.

Jevtíc, A., Valle, A. F., Aleny`a, G., Chance, G., Caleb-Solly, P., Dogramadzi, S., & Torras, C. (2018). Personalized robot assistant for support in dressing. *IEEE Transactions on Cognitive and Developmental Systems, 11*(3), 363–374.

L. Jiang, J.D. Hwang, C. Bhagavatula, R.L. Bras, J. Liang, J. Dodge, K. Sakaguchi, M. Forbes, J. Borchardt, S. Gabriel, Y. Tsvetkov, O. Etzioni, M. Sap, R. Rini, Y. Choi, Can machines learn morality? The delphi experiment, arXiv preprint arXiv:2110.07574 (2021).

Jirotka, M., & Stahl, B. C. (2020). The need for responsible technology. *Journal of Responsible Technology, 1*, Article 100002.

Johnson-Laird, P. N. (1989). *Mental Models*. The MIT Press.

Kagan, S. (1988). The additive fallacy. *Ethics, 99*(1), 5–31.

Kearns, M., & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics, 26*, 293–307.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies, 96*, 290–303.

Klein, G. A. (1993). A recognition-primed decision (rpd) model of rapid decision making. *Decision Making in Action: Models and Methods, 5*(4), 138–147.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456–460.

G. Klein, Expert intuition and naturalistic decision making, handbook of intuition research (2011) 69–78.

Lin, C. J., Lin, S. F., Wang, R. W., Sun, T. L., Chao, C. J., Feng, W. Y., & Tseng, F. Y. (2011). A skill-, rule-, and knowledge-based interaction design framework for web-based virtual reality training systems. *Key Engineering Materials, 450*, 564–567.

Lintern, G. (2010). A comparison of the decision ladder and the recognition-primed decision model. *Journal of Cognitive Engineering and Decision Making, 4*(4), 304–327.

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making, 14*(5), 331–352.

Liu, H., & Wang, L. (2021). Collision-free human-robot collaboration based on context awareness. *Robotics and Computer-Integrated Manufacturing, 67*, Article 101997.

M. L. Y. S. S. I. B. Y. A. R. d. M. V. T. B. B. H. C. A. Feng, N., R. Calinescu. (2024). Normative requirements operationalization with large language models. In *32nd IEEE International Requirements Engineering (to Appear)arXiv Preprint arXiv: 2404.12335.)*.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), Article 2053951716679679.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.

Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. *Decision Making in Action: Models and Methods, 1*, 3–20.

Parnell, K. J., Wynne, R. A., Plant, K. L., Banks, V. A., Griffin, G., Thomas, & Stanton, N. A. (2022). Pilot decision-making during a dual engine failure on take-off: Insights from three different decisionmaking models. *Human Factors and Ergonomics in Manufacturing & Service Industries, 32*(3), 268–285.

Parnell, K., Merriman, S., Getir Yaman, S., Plant, K., & Calinescu, R. (2023). Resilient strategies for socially compliant autonomous assistive dressing robots. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (pp. 1–9).

Quintas, J., Martins, G. S., Santos, L., Menezes, P., & Dias, J. (2018). Toward a context-aware human–robot interaction framework based on cognitive development. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49*(1), 227–237.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, 3*, 257–266.

Rasmussen, J., & Jensen, A. (1974). Mental procedures in real-life tasks: A case study of electronic trouble shooting. *Ergonomics, 17*(3), 293–307.

A. Rawls, Theories of social justice (1971).

Sheridan, T. B. (2017). Musings on models and the genius of jens rasmussen. *Applied Ergonomics, 59*, 598–601.

Siegel, J., & Pappas, G. (2021). Morals, ethics, and the technology capabilities and limitations of automated and self-driving vehicles. *AI & Society*, 1–14.

Stilgoe, J., Owen, R., & Macnaghten, P. (2020). Developing a framework for responsible innovation. *The Ethics of Nanotechnology, Geoengineering, and Clean Energy* (pp. 347–359). Routledge.

Townsend, B. A., Plant, K. L., Hodge, V. J., Ashaolu, O., & Calinescu, R. (2023). Medical practitioner perspectives on ai in emergency triage. *Frontiers in Digital Health, 5*, Article 1297073.

Townsend, B., Paterson, C., Arvind, T. T., Nemirovsky, G., Calinescu, R., Cavalcanti, A., Habli, I., & Thomas, A. (2022). From pluralistic normative principles to autonomous-agent rules. *Minds and Machines, 32*(4), 683–715. https://doi.org/10.1007/s11023-022-09614-w

Vallor, S. (2020). Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century. *Machine Ethics and Robot Ethics* (pp. 137–154). Routledge.

Vicente, K. J. (1999). *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. CRC press.

Ward, N. J. (2000). Automation of task processes: An example of intelligent transportation systems. *Human Factors and Ergonomics in Manufacturing & Service Industries, 10*(4), 395–408.

L. Weidinger, K. McKee, Veil of ignorance applied to selecting principles for AI (2022).

Weiss, A., & Spiel, K. (2022). Robots beyond science fiction: Mutual learning in human–robot interaction on the way to participatory approaches. *AI & Society, 37* (2), 501–515.

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

Williams, B. (1981). *Moral Luck: Philosophical Papers 1973-1980*. Cambridge University Press.

I. Young, Responsibility for justice oxford university press, New York (2011).