Minds and Machines
https://doi.org/10.1007/s11023-022-09614-w

From Pluralistic Normative Principles to Autonomous-Agent Rules

Beverley Townsend[1] · Colin Paterson[1] · T. T. Arvind[1] · Gabriel Nemirovsky[1] · Radu Calinescu[1] · Ana Cavalcanti[1] · Ibrahim Habli[1] · Alan Thomas[1]

Received: 29 April 2022 / Accepted: 14 October 2022

Abstract
With recent advancements in systems engineering and artificial intelligence, autonomous agents are increasingly being called upon to execute tasks that have normative relevance. These are tasks that directly—and potentially adversely—affect human well-being and demand of the agent a degree of normative-sensitivity and -compliance. Such norms and normative principles are typically of a social, legal, ethical, empathetic, or cultural ('SLEEC') nature. Whereas norms of this type are often framed in the abstract, or as high-level principles, addressing normative concerns in concrete applications of autonomous agents requires the refinement of normative principles into explicitly formulated practical rules. This paper develops a process for deriving specification rules from a set of high-level norms, thereby bridging the gap between normative principles and operational practice. This enables autonomous agents to select and execute the most normatively favourable action in the intended context premised on a range of underlying relevant normative principles. In the translation and reduction of normative principles to SLEEC rules, we present an iterative process that uncovers normative principles, addresses SLEEC concerns, identifies and resolves SLEEC conflicts, and generates both preliminary and complex normatively-relevant rules, thereby guiding the development of autonomous agents and better positioning them as normatively SLEEC-sensitive or SLEEC-compliant.

Keywords Normative principles · Social · Legal · Ethical · Empathetic and Cultural (SLEEC) norms · SLEEC rules · Autonomous agents

✉ Beverley Townsend
bev.townsend@york.ac.uk

Radu Calinescu
radu.calinescu@york.ac.uk

[1] University of York, York, England, UK

Published online: 29 October 2022

1 Introduction

Recent technological advances have allowed autonomous agents to become increasingly sophisticated. This promises great benefits to individual users and society alike. However, in the realisation of such promise it is important that these agents do not violate social, legal, ethical, empathetic, and cultural ('SLEEC') norms. A working definition for these interrelated norms is that they are 'the fundamental principles that govern the issues of how we should live and what we morally ought to do' (Driver, 2007, p. 32) or 'customary rules that govern behavior in groups and societies' (Bicchieri et al., 2018).

In this paper we will distinguish a fundamental set of principles from a much broader range of associated norms which we will call 'evaluative standards' (McKeever & Ridge, 2006, pp. 9–11). This distinction may be applied within each of the domains with which we are concerned: social, legal, ethical. In the specific contexts of application which are our primary concern in this paper, we think it is important to map out how the highest-level principles are related to context specific evaluative standards. These explicitly formulated evaluative standards, which may loosely be called 'rules', require not just the refinement but in some cases the adjustment of the principles which underlie them. We will call these local, context specific, evaluative standards, 'SLEEC rules'.

This paper seeks to demonstrate how we might derive particular, encoded, specifications of SLEEC rules for a given autonomous-agent task from a set of high-level principles. These evaluative standards, expressed as SLEEC rules, complement the functional requirements expected to be met by the agent, and support the use of techniques that can provide evidence that the agent's decisions and actions are SLEEC-sensitive or SLEEC-compliant.

Our aim is, by fully specifying this process, further to progress the issue of the trustworthiness of autonomous systems as they are put to use in ever more complex environments. Deriving specifications for trustworthy behaviours in robotic systems in complex environments is a challenging task (Menghi et al., 2019; Dennis et al., 2015; Miyazawa et al., 2016; Lindoso et al., 2021). The inclusion of properties which address SLEEC concerns into this assessment has been limited to date.

However, our proposal here builds on the intellectual effort and research on machine ethics (Allen et al., 2005, 2020; Moor, 2006; Anderson & Anderson, 2007; Winfield et al., 2019), defeasible reasoning (Thomas, 2011; Horty, 2012; Knoks, 2020), the dialogical and collaborative approach of 'doing' ethics by embedding normative values in agents (Stahl & Coeckelbergh, 2016),'ethical design' (or the process by which 'ethical values or principles are taken into account or embedded in the design process of a product, device or technology' (Yew, 2021), and 'value sensitive design' (integrating moral values in technology through design) (Manders-Huits & Van den Hoven, 2009; Van de Poel & Kroes, 2014; Umbrello, 2019; Umbrello & Van de Poel, 2021). In doing so, we offer a process whereby concrete rules for implementation may be derived from higher level principles by a process of specification (Richardson, 1997). We acknowledge that in practice the derivation of such rules is complex, and offer only a first

attempt at describing how SLEEC requirements underpinned by principles and specified by rules may be articulable in any given context. Our objective is to demonstrate a 'proof of concept' which has abstracted away from many complex issues of feasibility. There are, undeniably, technical limitations in adopting such an approach and its feasibility is grounded in the organisational realities faced by developers and users and complexities involved in the mapping process, for example.

By adopting a specificationist approach, we emphasise that we do not view this process of deriving context specific standards from principles as a mechanical one only of 'top-down application'. The identification of both high-level principles and lower level evaluative standards can help us to identify what counts as a relevant normative concern in the first place. Conflicts can be identified in order to be addressed even in cases where they cannot be fully resolved. Principles, as specified for new contexts of application, may need to be revised in the process. 'Application' is never mechanical, always involves judgement, and can be transformative of our most fundamental commitments (Thomas, 2006, p. 284). Our aim is to pave the way for a general conception of a process that not only spans disciplines, but aligns, strengthens, and furthers the application of established ethical design methods. This process tackles the difficult requirements engineering problem of 'converting vague goals [...] into specific properties' (Zave, 1997, p. 316), and is aligned with the view that 'requirements elicitation should focus on requirements for acceptable behavior' (Cheng & Atlee, 2007, p. 294). As such, we envisage that its integration within existing requirements engineering methodologies (Bennaceur et al., 2019; Pohl, 2010) will support the elicitation of SLEEC rules alongside other categories of nonfunctional requirements (for example, dependability, performance, and security) that are essential for high-integrity autonomous agents (Calinescu, 2013). Like any other nonfunctional requirements, the SLEEC rules derived using our process may be used to guide the development, verification, deployment, operation, and maintenance of autonomous agents so that they may be considered to be sensitive to such considerations or, in some crucial legal instances, compliant with them. The rules can be used to augment and complement the functional specifications of what the agent ought to do and in what order. Further, they place much needed constraints on resilience mechanisms by controlling and limiting the degrees of freedom of the agent.

The three central aims of this paper are as follows. First, we identify the high-level principles that are relevant to the development of SLEEC autonomous agents. Second, we offer and defend a rule-elicitation process that is conducive to the derivation of SLEEC rules from SLEEC principles. Third, we pragmatically demonstrate how this might be achieved using, as an example, an assisted dressing agent under development.

The paper is structured as follows. After describing related work, we introduce a robotic assisted-dressing case study used for illustration, as we set out the rule elicitation process. This comprises five stages that we describe in turn: (i) identifying high-level normative principles; (ii) mapping principles (and proxies) to agent capabilities and writing preliminary rules; (iii) identifying SLEEC concerns; (iv) identifying and resolving SLEEC conflicts; and (v) labelling, assessing their impact, and

developing complex rules by refining and extending preliminary rules by drawing on a non-monotonic logic to formalise the underlying inferences.

2 Background

Autonomous agents can no longer be considered what Wallach and Allen (2008) describe as 'ethically blind'. Agents perform a number of evaluative and personal functions. Such agents can potentially (and paradoxically) serve to either enhance or diminish human well-being. They can allow users to achieve more valuable ends and make more authentic choices or can serve to diminish authentic human choice (Formosa, 2021).

As agents move from instrumental tools to playing the roles of care givers and interactive agents (Breazeal, 2004), and as their prevalence increases, building and deploying SLEEC-sensitive agents becomes increasingly important. This is because agents of this type do not function within a normative vacuum, but exist within a specific social, legal, ethical, empathetic, and cultural milieu. As their roles, actions and choices expand, and because they exist in close proximity to users, often in personal spaces, and engage in personal (and, sometimes intimate) interactions with users, the appropriateness of their actions and choices involve considerations of a normative nature. This involves normative sensitivity and a level of normative decision-making - the implications of which are far-reaching.

The actions of an autonomous agent can involve moral, cultural, and social choices (in contradistinction to technical and operational choices that are not ethically and normatively charged). Actions and associated choices are often executed under non-ideal conditions, are often of significant moral risk, and have the power to directly affect human well-being. Such actions may involve the choice to treat one value as more important than another in a specific context: accuracy over fairness, privacy over accessibility, preventing harm over respecting the user's autonomy, or favouring individual-level interests in justice over group interests in safety and security. Consideration of a broader set of social, legal, ethical, empathetic, and cultural norms is required to determine what is appropriate within a domain involving judgements of compromise and trade-off.

We are committed to stakeholder engagement and dialogue as an integral part of our multi-stage process. However, we take this to be the basis of the legitimacy of the procedure and not a commitment to the view that continued dialogue is automatically a method of conflict resolution. Continued dialogue can expose deeper disagreement than was initially supposed. Our view, expressed in Thomas (2006, pp. 283–284) is that dialogue is integral to the definition of a 'problem situation'.[1] In such a situation, areas of agreement are identified in order to frame continued points of disagreement. Whether continued dialogue will resolve or sharpen such disagreements is not, in our view, something to be resolved a priori. The process,

[1] In contextualist terms, an overlapping consensus describes a problem situation, not a particular solution to any such problem (Thomas, 2006, p. 284).

however, confers legitimacy on the output. It would, indeed, be implausible to argue that handing off complex moral issues to a machine would lead to their resolution. Conflicting points of view would simply turn the focal point of their disagreement from the first order issue to the legitimacy of the machine and its outputs. Our more modest goal is to have identified both our agreements and disagreements with the hope that dialogue will be able to resolve some, if not all, hard moral conflicts. On any view there are some deep moral conflicts that we may simply have to live with.

One concept on our list requires further explanation: social norms, legal norms and ethical norms form complementary, overlapping domains of norms. Empathy, however, serves to identify a core human capacity that underpins our ability to reason across all of these domains. We have highlighted empathy as a generic term for our ability to be sensitive to whether things go well or badly for other human beings. We do not envisage, nor do we anticipate, building such a capacity directly into an autonomous system depending in the way that empathy does on the imagination. We take empathetic understanding to be important to those who design, program, and implement such systems in contexts where their behaviour does, indeed, either promote or inhibit human well-being.

Not only is the expectation that the activities and outputs of autonomous agents be compliant with SLEEC norms, follow the Asilomar AI Principles (Future of Life Institute, 2017) that 'highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values through their operation', and be 'compatible with ideals of human dignity, rights, freedoms and cultural diversity', but it is necessary to be able to trust that they do (Dennis et al., 2016). The SLEEC rules themselves can become evidence to support an argument that the resulting agent can be trusted to perform in a manner that is aligned with expected behaviours.

We believe that the work presented in this paper is important for two reasons. First, it allows for augmentation of autonomous agents with the ability to carry out decision-making and actions that not only meet technical requirements but are sensitive to social, legal, ethical, empathetic and cultural norms.

Second, faced with SLEEC concerns and conflicts, it offers a method to develop a bank of default, defeasible, rules which attempt to list specific defeaters. By a 'defeater' we mean a consideration that enables something to be a reason, or a consideration that rebuts a reason by supplying a contrary reason (Dancy, 2004). A stronger notion of undercutting defeat can be captured that undermines the relation between a reason and a putative conclusion (Pollock, 1987, p. 485). Once fundamental principles are identified, the well developed methods of non-monotonic logic can be used to represent the inferences from such principles to context specific evaluative standards (Horty, 2001). The process can then be repeated to capture inferences involving those standards themselves. The process itself will disclose potential conflicts and ways of resolving them (if available). Our overall aim is to defend a hybrid position (Asaro, 2006) that, in combination, is both 'top-down', by using principles as a point of departure, and 'bottom-up', where the tasks of the agent are constrained in accordance with a set of pre-defined rules underpinned by SLEEC norms. This process of rule elicitation allows SLEEC norms to provide a *pro tanto* reason for embedding a rule (or executing a course of action) within a given context.

Van de Poel (2009) and Latour and Venn (2017) emphasise the role of values in engineering design and the moral relevance of design decisions. The question of whether it is possible to create artificial full ethical agents is yet to be resolved (Tonkens, 2012, p. 139). This is not to say, however, that by treating autonomous agents as 'value-laden' or 'ethically- or normatively-sensitive' we mean to imply, or to conceive of them, as fully ethical or moral agents.[2] We suggest, only, that certain socio-technical agents can be made highly adaptive, interactive, and responsive to SLEEC concerns within a particular local context by the introduction of a set of defeasible SLEEC rules used to design and verify the behaviour of the agent.

In this paper we consider primarily the design of those agents that fall short of the ideal of being full ethical agents, but can nevertheless be designed in a way that implements the guidance of normative behaviour. Our approach is based on identifying, during the requirements engineering process and conducted with the input of stakeholders, a set of SLEEC rules to guide and restrict the behaviour of the agent. We refer to such autonomous agents interchangeably and broadly as the 'system', the 'robot', or the 'agent' as the context dictates in what follows. We refer more specifically to 'social' robots as autonomous or semi-autonomous systems that are designed to interact socially and communicate with humans and other robots, and to 'care' robots as those social robots designed to perform tasks 'related to physical or emotional care' (Goeldner et al., 2015, p. 115).

We base our design on a framework that distinguishes between principles, rules (or 'evaluative standards'), and actions. By 'principles', we mean high-level ideas such as 'dignity', 'autonomy', 'accountability', 'justice', and 'non-maleficence', which guide the conduct of moral agents generally and apply across a wide range of domains (Ross, 2002). Evaluative standards are derived from such principles and give them practical import by setting out guidance in relation to how a moral agent ought to behave in a particular context (Thomas, 2006). They are intended to shape the actions or course of conduct in which a moral agent engages, and the choices it makes, in response to a particular body of context specific knowledge (Henderson, 2002, p. 332).

Because principles are articulated at a high level of generality, they give rise to a plurality of normative principles both across and within the social, legal, ethical, and cultural contexts, each of which will in turn have a number of implications for the manner in which a moral agent should act. Table 1 illustrates this with reference to the norm of 'dignity' in the context of adult care. The process of generating SLEEC rules for guiding the design and operation of an autonomous agent involves ensuring that its actions in response to stimuli are modelled on those that a moral agent following the applicable principles would undertake in response to a similar body of information.

We have already noted that social, legal, ethical, and cultural normative principles significantly overlap. Any token action might involve considerations drawn

[2] On an account of the morality of artificial agents and on moral agents that can be involved in moral situations but do not necessarily exhibit free will, mental states, or emotions see Floridi and Sanders (2004).

Table 1 Norms, principles, and actions

Type	Social	Legal	Ethical	Empathetic	Cultural
Norm	Dignity				
Normative principle	The personal and social identity of users of care services should be respected and supported (National Institute for Health and Care Excellence, 2013, p. 13)	Users of care services should be treated in a way that ensures their privacy, autonomy, independence, and involvement in the community (Health and Social Care Act 2008 (Regulated Activities) Regulations 2014, reg 10)	Users of care services should be able to choose the care and support they receive (Skills for Care, 2013, p. 5)	Users of care services should be treated with compassion and enabled to engage in meaningful activities that use the skills and capacity they have (The Health Foundation, 2016)	Users of care services should be helped to maintain religiously mandated lifestyles (De Voogd et al., 2021)
Rule	Directives for execution by the agent that produce actions conforming to relevant principles				
Example action	Address the user by their preferred title, such as Dr., Professor, Rabbi, and so on	Protect the confidentiality of sensitive personal information pertaining to the user in interactions with others unless the user permits disclosure	Allow the user to do things they have expressed a preference for doing themselves such as making their own tea	Arrange facilities in a way that maximises personal mobility, for example by leaving doors open	Incorporate assistance in wearing religious vestments into care routines

from some, or all, of these domains. With an ecumenical aim, we have not tried to regiment sharp boundaries between these categories of norms—as one might do, for example, by treating culturally specific norms as a sub-set of social norms. Our purpose in identifying them as separate loci of concern is, rather, to model the full breadth of concerns and expectations to which autonomous agents operating in a social context must be sensitive. In practice, as Table 1 demonstrates, principles interact and inform each other. We have taken as a representative example the legal understanding of dignity as encompassing respect for a person's 'physical and mental integrity' (European Charter of Fundamental Rights, Article 3(1)) which has implications for not only the legal but also the ethical and empathetic normative principles to which the norm gives rise.

3 Robot Assisted Dressing : A Use Case

Developments in machine learning and control engineering promise a world in which robots are able to provide care and support for individuals in their daily lives (Jevtić et al., 2018; Zhang et al., 2019; Coşar et al., 2020). While a human carer may still be required, robotic autonomous systems may allow for increased reach, enhance existing activities, and enable greater multitasking. We consider the example of an agent that aids a user in dressing, an activity that traditionally involves care professionals. The system is assumed to be deployed within the user's home and, as such, aids in maintaining the user's independence, as well as allowing the re-allocation of resources in the care system. Whilst the primary role of the agent is dressing, a secondary function, to monitor the well being of the user, is also expected. This is an additional activity that a human carer would undertake quite naturally, even if it is not their primary role.

Figure 1 indicates the context in which the proposed agent is expected to work, connecting to the home automation system as well as a remote support unit. To carry out the dressing objective, the agent is equipped with moving actuators able to pickup and manipulate the clothing in close proximity to the user as well as multiple cameras that capture video imagery to determine user pose and limb trajectory. In addition the agent has voice synthesis and recognition to interpret verbal commands and communicate progress to the user. Communication with the user is also possible using a touch screen mounted on the robot. The audio-visual components may also be leveraged to monitor user well being through machine-learning components that detect distress in speech patterns as well as facial expressions. Finally, the user wears a smart watch that provides biometric information and has the capability to detect falls.

In completing their tasks human carers must balance concerns for ethical, social, and legal norms by drawing on an underlying capacity for empathetic understanding. Asking for permission before proceeding, closing the curtains before dressing the person, and making sure the temperature is comfortable are natural considerations of a human carer when tasked with dressing a person. These considerations are constitutive of, and implied in, the job of caring, and even more generally of most human-to-human interactions. Not only does this allow tasks to be completed more

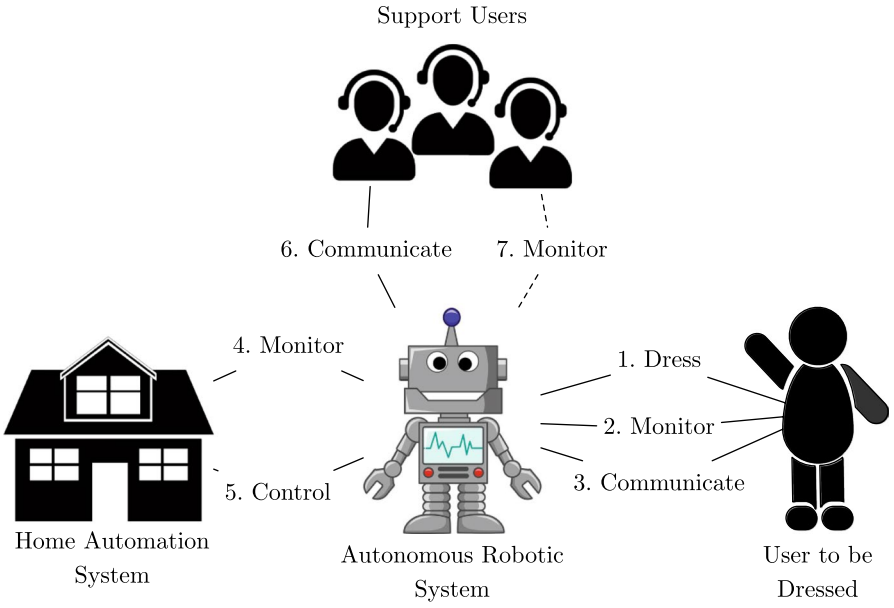

Fig. 1 Robotic assisted dressing application. An autonomous robotic system is used to dress its end user (1), while monitoring their well-being (2). The system may communicate with the user (3) receiving instructions for action and providing information and prompts as appropriate. The autonomous robotic system is additionally able to monitor the status of the environment (4) and control the home automation system (5). An assistive-care support team may be contacted where external human input is necessary (6) and the team may periodically monitor the status of the mechanical system (7)

effectively in the long run, but it enhances trust. This highlights the need for SLEEC concerns in agents tasked with caring for its users, to various degrees depending on the agent's scope of care. In order to maintain trust and confidence that an agent is functioning well, and will continue to do so, addressing SLEEC concerns is essential.

Given the limited capacity of agents to take advantage of machine learning at this time, and the inherent difficulties of machines in 'resolving' complex and highly context-dependent ethical dilemmas and in 'applying' legal theory and reasoning, it is human agents that derive the set of programmable rules in the process presented here. We do not view this as a drawback to our approach. Given the current state not only of the development of autonomous systems but also their uptake and use, end-users are likely to require assurance that these systems are regulated by moral principles before trusting elderly people and children to their care. These rules are integral to the operation of the SLEEC agent and are formulated following the process outlined below. It is instantiated through a dialogical process of collaborative engagement, and in deliberative and meaningful consultation, with stakeholders—that is, users, domain experts (such as carers and health practitioners), user advocacy groups, developers, designers, ethicists, philosophers, lawyers, community leaders, or members of the public, and so on. Consultation with stakeholders plays an important role in the direction taken to credibly align the rules with the norms of a diverse

base of stakeholders, including end users (Umbrello, 2018). While participatory design as a concept is not new, it has gained in popularity recently in AI technology development (Zytko et al., 2022; Hossain & Ishtiaque Ahmed, 2021). Such participation can serve to mitigate the adverse implications of autonomous agents on society and on vulnerable and marginalised populations (Zytko et al., 2022). Indeed, the suggested dialogical and participatory approach of involving stakeholders to embed normative principles in rule-writing forms an integral part of responsible research and innovation (Stahl & Coeckelbergh, 2016).

However, deriving rules from normative principles is not easily done: it is a task premised on human intuition, understanding, imagination and commonsense. A plethora of reasons might exist that plausibly count for why we ought to value one reason or outcome above another—and, equally, many reasons why we ought not to. And while we require shared human experience in the identification of normatively-relevant considerations and concerns, even this is problematic, as humans, themselves, are not always and necessarily legally-, ethically-, socially-, or culturally-competent or aware, and often disagree over fundamental principles and how they might be applied in practice. That said, the field of ethics has developed over centuries and has a cumulative tradition of enquiry into the fundamental principles of ethics. The status of the field of ethics as a science is controversial-Oxford philosopher Derek Parfit described it as a science that as in its infancy (Parfit, 1984, p. 154). But, in fact, moral philosophers have converged on a relatively narrow set of candidate fundamental principles (McKeever & Ridge, 2006, p. 194).

More challenging is deriving from these principles the evaluative standards that informed stakeholders and end users would agree are relevant to decisions in specific contexts. Exactly how broad (or, how many) or how narrow (or, how few) rules should be elicited might depend, for example, on the complexity and diversity of the determination, the system's capabilities and application, and the context. It is for this reason that we suggest that the discourse and activity of rule derivation should be inclusive, deliberative, and broad, drawing from a range of expertise and interested stakeholders. It may reasonably be anticipated that in certain circumstances agreement may simply not be reached. However, even where clear consensus cannot be secured, this should not be seen as a failure. Rather, it should be viewed as a way of seeking to promote better understanding of the complexities involved, of addressing concerns, and ultimately in finding legitimate and thoughtful paths to providing rule-based solutions. The identification of areas of convergence and agreement can be taken as a basis for framing points of disagreement in the hope that the latter may be overcome by further reflection. Alternatively, in instances of irreconcilable disagreement or conflict, the task lies in identifying such cases and formulating policies for how to proceed in the light of their irresolubility (Thomas, 2006, 284). The most sophisticated development and application of models of defeasible reasoning avowedly offer multiple formulations, some of which permit irresoluble conflicts or dilemmas and some of which do not (Horty, 2012).

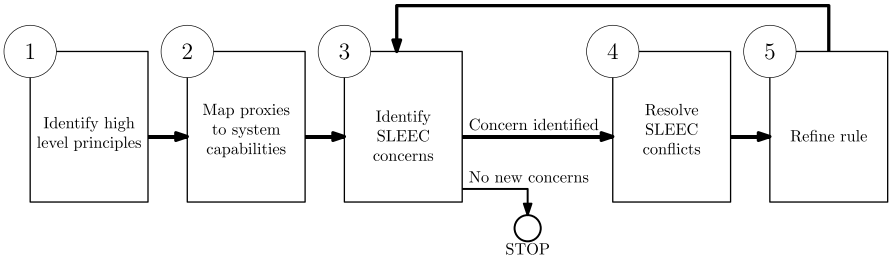

Fig. 2 Rule Elicitation Process

4 Rule Elicitation Process

Our contention, then, is that SLEEC rules are elicited from the collected insight of stakeholders within each of the SLEEC domains. The five-stage process shown in Fig. 2, and described in this section allows these stakeholders to refine high-level principles into rules that inform the design and operation of the agent. Bennaceur et al. (2019) identify four main activities within requirements engineering: elicitation, modelling and analysis, assurance, and management and evaluation. Our process belongs to the first category of requirements engineering activities, and is intended to be integrated within existing requirements engineering methodologies (Pohl, 2010) to allow SLEEC concerns to be captured for appropriate system development. Starting with principles and system capabilities, we progress through an iterative process in which SLEEC rules are shaped by SLEEC concerns and refined through the identification and writing of rule defeaters.

The first stage of the process identifies the high-level norms and principles that are relevant to the evaluative standards that agents actually apply in any given context. A plurality of normative principles are obtained from various sources including the articulation and listing of such fundamental principles in the history of moral philosophy. For our eirenic purposes we focus on the pluralist and non-prioritist (the plurality of principles are not internally ranked) view of W. D. Ross. In the second stage these identified principles are used to derive contextually meaningful proxies and placeholders (described hereunder in Sect. 4.2) that are mapped to the agent capabilities. This allows for the identification or flagging of primary 'touch points': areas where, given the agent capabilities and the operating context, preliminary rules involving a principle or proxy can be written. Each of these preliminary rules is examined in stage 3 to identify broader SLEEC concerns. Identification is achieved through consultation with stakeholders and domain experts. In stage 4 we identify and seek to resolve any SLEEC conflicts, which arise when two seemingly competing SLEEC concerns are identified (described in more detail in Sects. 4.3 and 4.4 hereunder). In stage 5 we extend and refine each of the preliminary rules, with the aim of addressing and resolving SLEEC conflicts, via the identification of relevant defeaters. As a newly refined rule may give rise to novel and different SLEEC concerns, the rule is passed back to stage 3 where any further SLEEC concerns (associated with the new rule) are identified. In this way, the process continues until all

identified SLEEC concerns and conflicts are addressed (and resolved) within a now complex rule, consisting of a preliminary rule hedged with multiple defeaters. When no further concerns are raised we proceed to the next preliminary rule. On completion of the assessment of all preliminary rules the process terminates.

In the following sections we describe in turn each stage in our process, and provide examples of application in practice using the use case.

4.1 Stage 1: Identifying Norms and Normative Principles

The first stage of the process consists of identifying high-level norms and normative principles. We consider principles that inform the system design in light of the agent's capabilities and the operating context into which the agent will be deployed. In this way we derive clear links (and dependencies) between the principles, the capabilities, and the limitations of the agent. At this stage we are not concerned with the specifics of the case, rather we wish to capture how the context shapes our understanding of, and the relevance of, the normative principles. We believe that the relative importance (or salience) of a principle, and the way in which a principle supports different evaluative standards, is a function of the context into which the agent is to be deployed.

The local context and the capabilities drive the rule elicitation process. The local context can lead to different conclusions being drawn about how a principle is relevant, qualified, and implemented in practice. The context also tells us something about the appropriateness of a principle. So, for example, fairness or privacy may depend on the context of a relationship, the event, or the conversation, and the time or place within which the principle and rule is applied. This speaks to 'contextual integrity'—or—that (epistemic) rules should be applied in appropriate ways (Nissenbaum, 2014). Likewise, principles may differ in salience and application. For instance, the application of a principle may be different for an agent deployed in a healthcare context to that of one used to approve loans.

In an attempt to create an agent that can be considered SLEEC sensitive, it is necessary to draw on a set of non-exhaustive principles, that encode our values including, *inter alia*, Ross (2002)'s theory of *prima facie* duties, which we think can be expanded to incorporate the concerns of an ethics of care that holds that moral action centres on interpersonal relationships and care as a virtue.[3]

Moral philosophy has a long tradition of assessing such theories, whether they are in competition with each other or are complementary, and assessing how long the list of ultimate principles ought to be. Resolving this issue goes beyond the scope of this paper, but, in any case, we can ensure maximally comprehensive

[3] Whether or not Ross's plural principles are reducible to a smaller set, or to just one principle, for example the monism of the rule consequentialist tradition or Kant's categorical imperative (in its Principle of Humanity formulation) remains a matter of scholarly controversy. For the first (consequentialist) claim, see Hooker (1996). Hooker's claim is contested by Stratton-Lake (1997) and by Thomas (2000). For the second (Kantian) claim see Audi (2004). Audi's claim is contested by Hurka, see (Hurka, 2007, pp. 64-72).

coverage by taking Ross's disjoined list of seven principles (fidelity, reparation, gratitude, non-maleficence, benevolence, self-improvement, justice) as paradigmatic (Ross, 2002). If another moral philosopher succeeds in shortening this list by reducing some of these principles to others (for example, the utilitarian derives them all from benevolence), then so be it. That project is ancillary to ours, which seeks to maximise the breadth of coverage of this candidate list. Construing the items on the list as identifying the right making features of actions-pro tanto reasons in our sense-immediately connects this pluralism to the project of formalising ethical reasoning as non-monotonic reasoning from a background context formed by such *pro tanto* principles (Thomas, 2011).

We also expand our list by supplementing it with further principles derived from ethics- and rights-based instruments, guidance frameworks, and sources of social and cultural values. This includes legal, ethical theory, social and cultural instruments, standards, professional codes of conduct, protocols, and guidelines (OECD.org, 2022; Yeung, 2020; UNESCO, 2021; BS8611, 2016). Against a plethora of recently published AI ethics documents and instruments, emergent core themes and principles can be noted (Jobin et al., 2019). The EU High-Level Expert Group on AI in the 'Ethics Guidelines for Trustworthy AI' has formulated, for example, a broad range of ethical principles and values to draw upon and incorporate into autonomous agent design and deployment (European Commission, 2019). In terms of these guidelines, trustworthy AI systems (and agents) should embody three central pillars: that is, be lawful, ethical, and technically robust (European Commission, 2019). Aligned with these pillars are key requirements for what is considered trustworthy AI, namely, human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; environmental and societal well-being; and accountability. Requirements that are to be evaluated and addressed throughout the agent's life-cycle.

Ethics documents typically share many of the principles centred around the four core bioethics themes of beneficence; non-maleficence; autonomy; and justice; a variation on Ross's list which drops some of its elements and adds autonomy as a distinct principle (Ross, 2002). It has been proposed by some philosophers that, in the case of autonomous systems, a plausible extension to the basic list will include the further theme of explicability identified by Floridi et al. (2018, 2021). Similarly, themes found in the newly developed AI regulatory and guidance landscape include the rights-based principles of a right to privacy, respect for human dignity, transparency and due process rights, rights to be informed, rights to self-determination and non-discrimination, and socio-economic, security, and welfare rights. The notions of bringing about good outcomes, reciprocal return for benefits received, and the correction of previous wrongs all appear on Ross's list of basic moral principles but have their counterparts in legal reasoning and broader social norms. We extend the enquiry to a broader philosophical discourse beyond those principles set out in institutionalised AI ethics guidelines and sector-specific, industry self-regulatory frameworks. An important supplementary part of ethics considers the agent tasked with applying principles: the domain of the virtues, such as patience, care, kindness, and tolerance. These themes have underpinned much of the developing corpus of ethics of care and

virtue ethics and on empathetic norms that encourage pro-social behaviours such as helping, and exercising compassion and co-operation (Olderbak et al., 2014).

Importantly, in offering such a process, we do not wish to suggest that the complexity or depth and breadth of a moral principle is capable of simply being captured and reduced to one or more encoded rules. Only, we suggest, that certain rules can be written that provide practical substance to certain key features embodied within the principle. Thus, a set of rules written about privacy does not necessarily capture the full extent of what the notion of respect for privacy may entail. What, however, the rules seek to do is through a process of concretisation, to materialise certain specific key features of privacy, and express them in terms of social, legal, ethical, empathetic, or cultural requirements. The identification and delimitation of the principle supports its protection by rendering it, within the rule, dogmatically manageable and practically operational. Rules and instructions that then enable and facilitate what minimally may be understood to be privacy-preserving within a specific context.

Social robotic agents demonstrate a degree of sociability and emotional perception, by, *inter alia*, their engagement in high-level interactive dialogue, responsiveness to social cues, gesturing, mimicking human social behaviour, and voice recognition (Darling, 2016; Formosa, 2021). This serves not only to facilitate the human-robot interface but also to promote their self-maintenance, learning, and decision-making capacity (Breazeal, 2003). However, while the agent might have the capability to reflect human behaviour and emotion and to exhibit the external or '"outward" aspect of care' (Yew, 2021), it is devoid of the capacity for empathy in any 'real' sense. Empathetic norms and concomitant pro-social behaviours and responses with regard to what is considered by humans 'to care', 'to be caring', 'to act with compassion', or 'to be cooperative', based on what is understood in human experience to be empathetic within a given context can be encoded as empathetic rules. Norms of explicability and transparency may, for example, make it impermissible to create autonomous systems that 'mimic' care from the point of view of an end user (Pasquale, 2020, pp. 9,80).

Additionally, we can derive rules from various cultural and social normative sources. Cultural norms conceptualise the values that underlie a cultural unit within society. They are the shared meanings assigned by members of the culture to things and persons around them, and the shared expectations that guide the behaviour of people within the cultural group (Smith et al., 2002). A cultural norm may indicate the culturally appropriate and acceptable way of dressing or addressing or greeting another in the context of the practices of that culture. Sources of guidance include cultural texts, religious doctrines, and the testimony of members and leaders of the cultural community. Similarly, social norms are those of a social nature that are shared among members of a social group that direct what is considered by a group of people within society to be a socially acceptable way of living (Legros & Cislaghi, 2020). They would include, for example, socially appropriate responses and dialogue (not swearing) and indicate what might be considered polite and acceptable behaviour in interactions with others (the way we treat our elders, for example). They have emerged as salient

Table 2 SLEEC sources and classification

Source	Social	Legal	Ethical	Empathetic	Cultural
Rights-based instruments		x	x		
International conventions		x	x		
Laws and regulations		x			
Ethics guidance documents		x	x		
Ethical theories	x	x	x	x	x
Codes of conduct, standards, and protocols		x	x		
Religious doctrines, customs, cultural and social texts	x			x	x
Community engagement	x	x	x	x	x
Input of domain experts	x	x	x	x	x
Members of the public	x	x	x	x	x

points of coordination between individuals and are therefore reflected in expectations about the stable underpinnings of social life in the future.

As indicated by the summary in Table 2 SLEEC norms often intersect and overlap. While certain normative principles, such as those derived from legal norms, may be well-established and prescribed—how personal data are processed, or what may be considered a rights-violation, for example—other norms, such as the dressing of the user in what society deems to be a presentable manner, albethey equally compelling, are less demanding and often not as clearly defined. Norms may also span SLEEC labels, so what it means to 'respect privacy' may be positioned differently with very different consequences and sanctions within a social, legal, ethical, or cultural setting. Where an infringement in one context (social or cultural, for instance) may result in embarrassment or shame, the infringement of the normative principle expressed in another context (legal), may result in non-compliance with a mandatory statutory requirement or civil action. SLEEC norms inform what we describe as the 'normative core' of a pluralistic principled approach to SLEEC system design and will have a direct effect on the development of SLEEC-sensitive autonomous agents, generally, and assisted dressing autonomous agents, more specifically.

A single-principle view, such as 'maximising utility' or 'maximising the intrinsic value of outcomes', can be justified either as an heuristic or an approximation in predicting the truth about micro-motivations. However, when applied (more generally) to the aggregate, we have challenged the adoption of a single-principle view, and suggest that, save for the very simplest of use cases, the process should be wider-reaching and highly adaptive to facilitate a complex, dynamic, and resilient context. The process we describe is one based on a plural principle approach (Ross, 2002). We have noted that some moral philosophers want to go further and ground this plurality on a narrower set of principles-sometimes merely one. We have not entered this debate in this paper; but we noted that the challenge is to the status of our principles as fundamental. For our critics, these principles are not fundamental; rather,

they are derived. But that does not mean that they deny that these plural principles are true-the locus of disagreement is elsewhere. That justifies our pragmatic and eirenic approach where we opt, instead, for Ross's original list of principles without seeking further to reduce the list in any way. This, we suggest, focuses the argument where it ought to be focused: on the difficult task of making the process of rule formulation more inclusive, adaptive, and explanatory.

We are, however, not committed to the view that norms or principles necessarily play the same role or have the same strength and salience in every context, rather and only, that there is a pluralism of norms (and reasons), supporting or refuting an action (or favouring one action above another) within a particular context. These are norms (and reasons) that will provide a scaffold for building a case for explicitly selecting (either by justifying or refuting) a course of action underpinned by a set of rules.

4.1.1 Application

In the assisted dressing use case, for instance, we can describe 'beneficence' as a relevant and applicable principle in that the agent should aim to benefit the user; 'non-maleficence' in that the agent should avoid harming the user—either psychologically or physically; 'autonomy' in that persons have the effective capacity to make decisions of their own that are of practical import to their lives, and that the agent should respect the wishes of the user and obtain consent (or assent) as the circumstances dictate; and 'privacy' in that the agent should not only not compromise the user's privacy and avoid spatial intrusions when executing the dressing activity by taking necessary precautions, but also safeguard their informational privacy rights by practising good and lawful data-stewardship measures.

Once we have established what are the normative principles that are relevant for the design of an agent, we proceed to the next stage, that of identifying proxies and agent capabilities, and mapping the former to the latter.

4.2 Stage 2: Mapping Principles to Agent Capabilities

Having completed the first stage of the process, we now have a list of the high-level normative principles that are applicable in the operating context in which the agent is to be deployed. At the end of this second stage, we will have a map connecting operationalisable principles to agent capabilities and a set of preliminary rules from which to derive our SLEEC rules.

This stage involves three steps: first, identifying proxies (or placeholders) for normative principles, second, establishing the relevant functional capabilities of the agent within the use case, and finally, mapping proxies to the capabilities. We assume here that the functional capabilities of the agent are known since, in line with the established requirements engineering practice, the SLEEC and other key nonfunctional requirements of the agent need to be devised alongside its functional requirements (Chung et al., 2000; Glinz, 2007).

Table 3 Components and capabilities of the robotic assisted dressing system

Component	Capability
Cameras	Estimate user pose, estimate limb trajectory, health assessment
Smart watch	Health assessment, fall detection
Audio speaker	Issue commands to the user, explain actions to the user
Voice recognition	Understand spoken requests from the user, obtain accent from the user
Touch screen	Accept commands from the user, obtain accent from the user, display support team member to user
Support Interface	Relay voice and video feeds from the user to the support team
Home automation interface	Monitor temperature in the house, turn light on/off, open/close curtains

First, normative principles in the abstract are not in the business of telling us about their practical application, their contextual appropriateness, or how to translate and operationalise them within a domain or sector. That is not their role which is, rather, to catalogue the right making features of actions. As Richardson suggests 'the crucial question is how ethical norms reach down to individual cases'(Richardson, 1990). For abstract normative principles to play their role they ought to ground usable evaluative standards. Such standards are narrower in their scope than principles and more obviously context dependent. If appropriate we will, in our method below, describe a 'proxy' (or a 'placeholder') for the normative principle. This is the actionable form that is used to represent the value of the underlying principle in the guise of a standard. For example, granting consent or obtaining assent can act as a proxy for the rights to autonomy and self-determination, and the ability to request and access information regarding the decisions and inferences made by the agent may be the specific and actionable form of more general rights to be informed and to transparency.

Second, we are only able to derive rules for 'execution' by an autonomous agent where the requisite system capabilities exist to facilitate such execution. Simply put, we cannot write rules around emotion- or facial-recognition, for example, if the agent is incapable or not equipped to perform such functionality. It is therefore necessary at this stage to enumerate the capabilities of the agent and the components that enable such capabilities. We note, however, that the process of SLEEC rule elicitation may itself drive the need for the addition of functional capabilities in order to allow the agent to meet such requirements—a common scenario of interdependence between functional and nonfunctional requirements (Chung et al., 2000).

We expose the agent capabilities through a conversation with roboticists who understand the hardware and software to be deployed in the agent. These capabilities represent the interface between the agent and the physical world, and it is through the application of these capabilities that the service is delivered and that infringements of principles (described in detail hereunder as 'SLEEC concerns') may arise.

A list of the components and their capabilities appropriate and available to our assisted-dressing example is provided in Table 3.

Third, having identified the primary relevant principles and their proxies within a use case as well as the capabilities of the agent, we begin a mapping exercise in

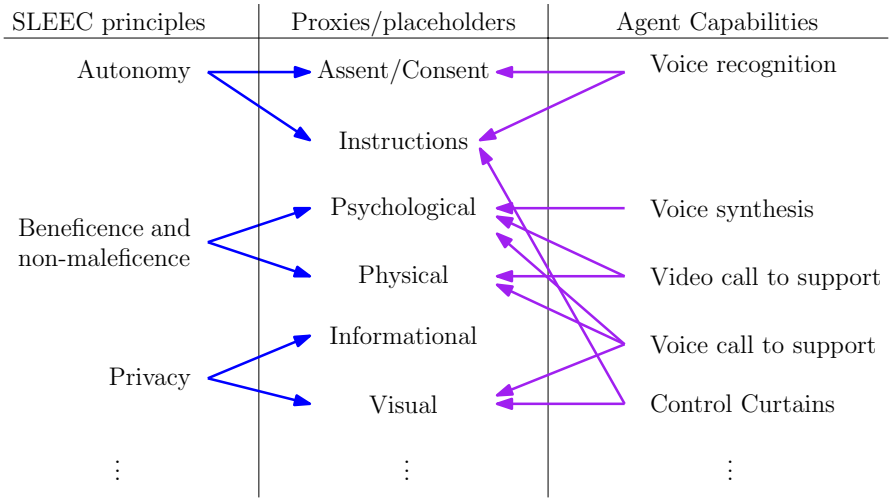

Fig. 3 A subset of the mappings between SLEEC principles and agent capabilities for our use case

which stakeholders consider how each capability may impact a normative principle (or proxy). This is a scoping exercise achieved through a discussion with stakeholders in a guided conversation in which we ask, 'what is the scope and extent of the principle?'; 'what is the essence of, and what makes up, the principle?'; 'what does the principle seek to protect or mean within this use case?'; and 'given the agent capabilities, how can such a principle (or proxy) be implemented in practice?'.

Through an understanding of the capabilities, we are able to consider unintended consequences that may compromise a principle. So, for example, a motion sensor implemented using infrared or ultrasonic sensors may have limited privacy implications, whereas a motion sensor implemented using a camera may give rise to increased privacy concerns.

4.2.1 Application

To demonstrate the application of this stage of our process, we consider again the use case presented in Sect. 3. Starting with the three high-level moral principles identified in stage 1 (respect for autonomy, privacy, and non-maleficence) stakeholders identify six proxies or placeholders. Consent or assent is the proxy for respecting a user's autonomy. Respecting autonomy is to ensure that the user maintains an appropriate level of control, so the act of giving instructions is expressed as a placeholder for the principle of autonomy. Non-maleficence comprises two placeholders concerning preventing harm to the user's psychological and to the user's physical well-being. Finally, privacy consists of safeguarding intrusion into the user's private space when they are, for example, partially clothed or undressed, and upholding informational privacy or practising sound data-protection practices.

The developers of the agent list the following amongst the system capabilities: voice recognition, voice synthesis, the ability to call support using audio or video

communications, and the ability to open or close the curtains. We are able to construct a map from proxies to capabilities defined for the system as shown in Fig. 3. For example, we may establish that voice recognition only exists so that the user can give instructions to the agent. Hence, and in this way, this capability allows for instructions and user assent to be received and therefore autonomy to be respected. Voice synthesis is included such that reassurance can be given to the user and, hence, a link is made by the healthcare professionals to the psychological health of the user.

Capabilities may be linked to multiple proxies. This is the case for voice-only support calls. Here the intention is that the agent will contact support when the user is in psychological or physical distress. In addition, the choice to use voice-only is motivated by the use case and the knowledge that a user may be in a state of undress and privacy should be maintained.

Figure 3 shows the results of the first phase of such an exercise for a subset of the proxies and capabilities of our assisted dressing applications. Here the links indicate that the agent has a capability that may affect a normative principle or its proxy and that a preliminary rule can be written about it. For example, we note that there are links for psychological health to three activities: calling support, asking of permission, and detecting distress. Similarly, we note that the action of asking permission requires a consideration of user assent or consent, dignity, and the psychological health of the user.

Further questioning the intention of the capabilities will lead to a set of preliminary rules that act as an input to the next stage of the process. Some examples of such rules derived from our map may be:

- *When the user tells the robot to open the curtains then the robot should open the curtains.*
- *When the robot cannot find a garment then the robot should inform the user.*
- *When the user is distressed and the user is fully dressed then support should be called using video.*

We derive a list of preliminary rules. Each preliminary rule is then considered, in turn, as input to the subsequent stages. We consider, as an example, the first preliminary rule described above.

4.3 Stage 3: Identifying SLEEC Concerns

Once we have a preliminary rule generated in stage 2, we proceed to stage 3, which involves identifying and considering SLEEC concerns. These are points (of impact) within the system-process that directly (and potentially, adversely) affect a SLEEC norm so that they are a cause of concern. We illustrate this issue using a preliminary rule from the use case. Namely, a rule shown below has been generated based on the need to respect user autonomy by directing the agent to follow a user's instruction.

> *When the user tells the robot to open the curtains, then the robot should open the curtains.*

However, given the context of assisted dressing, by opening the curtains privacy may be compromised, which is a cause for concern. Thus, this preliminary rule leads to the identification of privacy as a SLEEC concern. We do not seek to eliminate the preliminary rule, but rather to identify that a SLEEC concern exists, and to indicate that the rule should be extended to consider scenarios where following the rule puts privacy at risk. To assist in the identification of normative risk and SLEEC concerns, a range of tools can be deployed: including, impact assessments—be they data protection, fairness and bias, ethical or human-rights impact assessments. UNESCO has recently introduced the notion of an Ethical Impact Assessment in its Recommendations for the Ethics of Artificial Intelligence (UNESCO, 2021, pp. 50–53). This is one way to identify and assess the concerns and risks the agent poses to the user by the infringement upon one or more normative principles.

Legal concerns can be identified at this stage. Regulatory policy and standards applicable to the agent should also be ascertained. These are laws, specific regulations, policies, standards, and codes of conduct relevant to the field and include, more generally, adherence to any obligations of a legal or regulatory nature. Legislative and regulatory compliance can mandate certain activities, such as the strict adherence to stipulated data protection and transparency measures (see the EU General Data Protection Regulation 2016/679, for example, and the introduction of data protection measures), and the compliance with health and safety standards for robotics.

4.3.1 Application

An example of a(n extended) rule that takes the SLEEC concerns into account may require to collect only the minimum required personal information (data minimisation rules). Moreover, 'high-risk' systems may require stricter compliance duties. If 'emotion-recognition' systems are used to detect, for instance, whether a user is 'distressed', 'upset', or 'frustrated', this may trigger a legal duty regarding the disclosure requirement to inform the user that such a system is being used (see, for instance, the proposed EU AI Act) (European Commission, 2021). Thus, a preliminary rule written using such technology, would raise a SLEEC legal concern, and would need to be addressed, for example, by revising and extending the rule (as described in stage 5).

In this stage, it is not only instances of legal and ethical concerns that are identified. We also identify concerns of a social, empathetic, or cultural nature. In the context of the dressing robot, there are situations where the user's emotional state might give rise to concerns that necessitate an agent response of compassion, helpfulness, and cooperation requiring the generation of empathetic rules. Privacy may be challenged not only with respect to the user's informational privacy (and addressed by written SLEEC rules of a legal nature), but also with respect to the physical, psychological, and social dimension of privacy by intrusion into a user's personal space (and addressed by legal and social-cultural SLEEC rules) (Lutz et al., 2019). Examples of SLEEC concerns for our use case are provided in Table 4.

Once we have identified points of SLEEC concern, we ascertain whether or not any SLEEC conflicts have arisen in the next stage.

Table 4 Examples of SLEEC concerns in the robotic asssisted dressing system

SLEEC Concern	Description
Privacy	Limiting intrusion on the personal space of the user and ensuring privacy is protected; safeguarding health data, practising good data stewardship, and granting or restricting access to medical records
Respect for Autonomy	Granting and withdrawing of permissions, including consent and assent; ensuring the user maintains an appropriate level of control
Dignity	Understanding and accommodating the user's social and cultural sensitivities, respectful treatment
Explainability and transparency	Informing the user about system decision-making and any inferences made; providing justification for a course of action adopted
Beneficence	Maximising good outcomes
Non-maleficence	Minimising harm by ensuring safety and reducing the possibility of physical and psychological harm to the user

4.4 Stage 4: Identifying and Resolving SLEEC Conflicts

In stage 4 we identify and seek to resolve SLEEC conflicts where possible. The agent may both support and threaten different normative principles which in practice often requires trade-offs between different legitimate, yet conflicting, principles. Accessing personal data, for example, may improve the quality and efficiency of services, but compromise privacy and informational autonomy (for example, in the event of a security vulnerability). Similarly, increased automation, while a source of convenience, risks undermining autonomy and self-determination (Whittlestone et al., 2019). Other examples include: efficiency versus safety, predictive accuracy versus explainability, and autonomy versus beneficence. Given the right contextual factors, 'technologies might create tensions between any two (or more) of these values or norms—or even simultaneously threaten and enhance the same value in different ways' (Whittlestone et al., 2019).

A normative 'conflict' refers to the situation where actions A and B ought to be performed, but it is impossible to perform both (Horty, 2012, p. 65). We describe such actions as 'competing' or in 'tension' in the sense that certain decision contexts require resolution by negotiated justifiable trade-offs as either only one interest or value can be the most important in a given case, or a balance must be sought in establishing the 'sweet spot' where a position of compromise is favoured. 'Conflict' and 'tension', thus, speak of the ways in which the pursuit of one normative principle can resist or oppose another in a certain context (Horty, 2012).

4.4.1 Application

We suppose that in the deployment of an agent competing or conflicting normatively-relevant reasons arise that generate a conflict of obligations. For instance, we consider again the rule from our use case identified in stage 3 and reproduced below for convenience.

When the user tells the robot to open the curtains, then the robot should open the curtains.

In generating this rule, we have set up two potentially conflicting SLEEC norms, that is, respecting autonomy (by the agent following a user's instruction) and safeguarding privacy (by not following the user's instruction and ensuring the curtains remain closed).

We consider too the example of an agent that might be required to trade-off the principle of respect for autonomy against one underpinned by non-maleficence. For instance, a user may exercise their right to autonomy (by refusing that an action is performed, such as to take medicine, or get dressed as required by the agent) only to expose themselves to harm. Ethically-relevant reasons exist both for the system to respect the individual's autonomy and to uphold the principle of preventing harm (non-maleficence).

Both courses of action present reasons both to do something and not do something, underpinned by *prima facie* principles and obligations. Both have metaphorical weights and both have identifiable courses of action that rules can be written about and that are ethically indicated in the circumstances. We proceed then to resolve these conflicts as explained next.

4.4.2 Identifying Conflicts

We take the process of rule elicitation to be specificationist in this sense: a proposed principle is further specified by rules which articulate its implicit content. This introduces an ambiguity into the idea of completeness. We have made the idealised assumptions that the initial set of principles is 'complete'. However, the process of the further specification of the principles is, by its very nature, incomplete. Were that task completable, even at the limit, then we could build all possible specifications of the rule into the principle itself, producing an impossibly long, unusable (and unlearnable) conjunction of a principle with all possible circumstances to which it could be applied by specifying it. As John Horty has put it in a related context 'No satisfactory rule of this form has ever been displayed, and it is legitimate to doubt our ability even to formulate such fully-qualified rules with any degree of confidence, let alone learn these rules or reason with them' (Horty, 2012, p. 149). Our aim, for such context dependent reasoning, is to capture its indefinite character by explaining how an initial finite set of finite principles is indefinitely extensible (Thomas, 2011).

We take our view to be orthodox in its distinction between the principles that ground reason giving considerations in particular cases and that which the agent has, overall, most reason to do. The latter are often called, following Philippa Foot, 'verdictive' considerations 'all things considered' (Foot, 2003). There is a considerable literature on the relation between supporting reasons and the reasons that they support and resolving all forms of conflict is not our aim in this paper. Our more restricted task has been to list what we take to be the most plausible, plural, set of underlying principles governing the SLEEC domains. We have deliberately not tried to rank these principles inter se. All we think we need to say, for our purposes,

is that in any given context of application some reasons will present themselves as more important than others. Reasonable judgements can be made on that basis, even when comparison cannot involve commensuration on a single scale.

As discussed, determining a rule may require a degree of sacrifice in one norm in the pursuit of another. One way of resolving tension, we suggest, is to select the principle that is most important relative to a context. When this happens, we decide 'which of these values is more important (or, more precisely, [we] assess the importance of the marginal increments and decrements of these values that are at stake)' Scanlon (2003). We refer to this importance as the 'salience' of a norm. This is the evaluative quality of a norm that is particularly noticeable, important, or prominent, and which serves as the justification for a decision to introduce a rule. Based on the salience of a normative principle within a context we can establish a priority of normative principles and rules. We capture this from the considered judgement of the stakeholders.

4.4.3 Application

The withholding of assent, for example, as the proxy for respect for autonomy, would be grounds to justify the agent choice to not proceed with dressing a user. However, if the situation should change (by the introduction of new evidence, by the uncovering of a new SLEEC concern, or by obtaining a new instruction from the user), a different underlying normative principle may become relevant (such as in the prevention of harm) resulting in a rule to follow a new or revised course of action and associated choices.

We demonstrate next how to resolve conflicts and tensions by means of preliminary default rules and defeaters.

4.4.4 Deriving Defeaters

Having identified a preliminary rule in the previous stage, we now identify, via stakeholder engagement, the conditions in which a rule may be defeated.

Drawing on a defeasible reasoning framework, we establish the default rule (typically, the preliminary rule) with exceptions (dependent upon a use case and the use requirements) (Horty, 2001, 2012). Thus, we make such rules contingent upon the possibility of them being excluded or defeated by further specific reasons (known as 'defeaters'). This is done by considering the conditions under which the defeasible rule is invalid and setting out why and when those conditions do not hold. These defeaters address specific SLEEC concerns with the aim of resolving SLEEC conflicts such as those illustrated above—an aim, which we have stressed, may not be attainable in all cases.

This stage allows for the reasoner (in this instance, the stakeholders and rule writers) to draw plausible and tentative, but not infallible, conclusions that can subsequently be retracted based on further evidence. Moreover, it creates a mechanism of revising norms and rules in the face of the acquisition of new information (Reiter, 1980, 1988). This non-monotonic reasoning provides an efficient method of managing incomplete, dynamic information, where conclusions can be revised and

retracted as more evidence becomes available (McDermott & Doyle, 1980, p. 42). In this process, new SLEEC concerns and a changing context can lead to the withdrawal of previously established rules and the re-writing and extending of rules.

Thus, a rule can be defeated or overcome by exceptions (or 'hedges') (Horty, 2012). We follow the arguments of Väyrynen (2009) and Knoks (2020), who hold that moral principles and duties have 'hedges' (or built-in 'unless' clauses) which set out the conditions or circumstances under which the normative principle (or rule) does not apply. In practice, applying a hedge would mean that the system proceeds along a course of action underpinned by a preliminary, default rule, unless a hedge clause is triggered. We thereby account for as many exceptions (or deviations) from the default, preliminary rule as necessary.

Defeating conditions, represented by 'hedging clauses' are introduced to establish whether a rule is true by considering any counter-examples, any conditions under which a rule may not be valid, and by considering whether there are any reasons that may lead to the conclusion that the evidence supporting a rule might be invalid (Weinstock et al., 2013). Importantly, we cannot be sure that all possible defeaters have been identified within a context, only that a process is in place to anticipate and accommodate the finding of further evidence and for the creation of as many defeaters as may be necessary to accommodate this. We have noted Horty's argument that the goal is not ideally to complete the process with a 'bullet proof' principle that includes all its possible defeaters or hedging clauses and therefore makes any further process of specification irrelevant.[4] We refer also to Alan Turing's observation that 'It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances' (McKeever & Ridge, 1950, p. 452) which, we suggest, is pertinent to SLEEC rules.

4.4.5 Application

We refer again to the preliminary rule in the use case:

> *When the user tells the robot to open the curtains, then the robot should open the curtains.*

In this form, it favours autonomy. However, given the salience of privacy in the assisted-dressing context, we hedge it with the following 'unless' clause:

[4] We have not, in this paper, addressed the adjacent dispute between so-called particularist and generalists over the nature of moral judgement. We have noted that both Holton and Horty take themselves to be, in Holton's phrase, 'principled particularists' with Horty emphasising more strongly than Holton that any principled ethical understanding (such as the specificationism we have described here) depends on an ethical understanding that is not, and cannot be, principled (Horty, 2012, p. 163) (Holton, 2002). Similar arguments underpin the argument of Thomas (2011). However, the leading defenders of generalism-the view that ethical understanding can take the form of the grasp of a finite set of finite principles-are Sean McKeever and Mike Ridge in their book Principled Ethics: Generalism as a Regulative Ideal (McKeever & Ridge, 2006). They are committed to the view that ethical principles take this special form: they are material conditionals which, in their antecedent, quantify over all the relevant known defeaters. In that sense-the sense Horty disputes— this 'completes' the specification of the principle.

UNLESS the user is 'undressed', in which case the robot does not open the curtains and tells the user 'the curtains cannot be opened while you, the user, are undressed'.

The rule is now expressed as:

When the user tells the robot to open the curtains then the robot should open the curtains, UNLESS the user is 'undressed' in which case the robot does not open the curtains and tells the user 'the curtains cannot be opened while you, the user, are undressed.'

This now provides for a condition under which the preliminary rule will not hold, but will be defeated in the interest of safeguarding privacy. Further defeaters may be derived for this rule, for example, to deal with the scenario in which the lights are switched off at night, and therefore the curtains can be opened as requested without violating the user's privacy.

4.5 Stage 5: Labelling, Identifying Impact, and Re-assessing Complex Rules

We generate rules that capture the complexity of SLEEC concerns of autonomous agents. We have demonstrated how preliminary rules may be extended using defeaters creating 'complex rules', which comprise the preliminary rule together with single or multiple defeaters, as the use case may require. However, this is not the end of the process. In generating a complex rule, we need to re-assess it against any novel SLEEC concerns that may arise as a consequence of this new rule. This is done with due consideration of the impact the rule has on one or more principles. We next describe how a rule might be labelled and have its impact assessed, and then be re-evaluated in light of any anticipated or subsequent SLEEC concerns and conflicts.

4.5.1 Labelling and Accounting for Impact

Rules are labelled, according to their SLEEC type—as social, legal, ethical, or cultural. Rules written about a particular norm, such as dignity, for instance, may have different SLEEC types and depend upon the SLEEC concern the rule seeks to protect. So, as illustrated in Table 1, dignity may be concurrently a social, legal, ethical, empathetic, and cultural norm, but a legal rule written apropos dignity might look different and have different implications to one safeguarding a cultural dignity norm.

What is required in the process is not that SLEEC conflicts be avoided, but that the process anticipates such conflict within a context and has the ability to resolve conflicts in the face of one or more 'competing' normative concerns in so far as a ranking in terms of importance is possible. This is directly informed by the impact and salience such a rule may carry, which in turn informs the additional defeaters that may need to be written.

A rule, thus, informs a SLEEC concern positively, negatively, or in a manner that is neutral. This we refer to as its 'impact'. The impact of a rule assists in the management of conflicts, by the consideration of possible trade-offs and prioritisation, that is informed and accommodated through the generation of defeaters. We identify

impacts on a severity scale, that is, of high, moderate, or low severity. Identified impacts, such as those with the likelihood of causing serious harm, for instance, can be classified with 'high' severity and prioritised accordingly.

4.5.2 Re-assessing SLEEC Concerns

Finally, rules are re-evaluated and tested against new or anticipated SLEEC concerns and conflicts. Accordingly, rules are amended and new defeaters advanced, as required. On the strength of the information available, and as new information comes to light, so a rule, its priority (ordering or ranking), and its defeater(s) may be extended or changed. Thus, the adoption of a rule may need to be surrendered in the face of additional information and a changed context (Reiter, 1988). In this way the adaption and refinement of rules, through a process of iteration, can be better accommodated and aligned in a broad range of scenarios of use of a particular application.

We have explained the importance to our overall, specificationist, conception of practical reasoning that this process be indefinite. Not finite-that is the flawed ideal of seeking to 'complete' a principle by building all possible hedges into its formulation. Nor infinite-both principles and evaluative standards have to be learnable and usable by contingently limited agents like us. We anticipate that this process should practically continue until such time as it is sufficiently (and reasonably) obvious that the refinement of a rule (and any associated defeaters) has no further relevant specification or, as suggested by Weinstock et al. (2013), until such time as 'no increase in confidence will result from further developing the argument'. Reasoning must have a stopping operator, analogous to Richard Holton's 'That's it!' clause in the formulation of his own version of a 'principled particularism' (Holton, 2002) (Horty, 2012, p. 155 fn. 11)(Thomas, 2011).

4.5.3 Application

For the example, we have generated the following complex rule.

> *When the user tells the robot to open the curtains then the robot should open the curtains, UNLESS the user is 'undressed' in which case the robot does not open the curtains and tells the user 'the curtains cannot be opened while you, the user, are undressed'.*

This rule as it stands, while having a positive impact on privacy and explainability, may have a negative impact on the user's autonomy. Based on the view that privacy is salient, in this instance, we have traded autonomy off against privacy. However, and importantly, the enquiry is not complete. In this final stage of the process we need to recheck the rule against any novel SLEEC concerns that may arise as a consequence of the complex rule. So, in our example, by not following the user's instruction the user may become highly distressed or aggravated by this imposition on their autonomy, causing the user undue and unwanted psychological harm. This rule, thus, raises a new SLEEC concern: that of compromising the principle of preventing harm (or non-maleficence). We identify non-maleficence as a salient SLEEC concern in the context with a high-severity negative impact and so extend

the complex rule by writing a harm-avoidance rule. We now generate a further defeater:

> *When the user tells the robot to open the curtains then the robot should open the curtains, UNLESS the user is 'undressed' in which case the robot does not open the curtains and tells the user 'the curtains cannot be opened while you, the user, are undressed,' UNLESS the user is 'highly distressed' in which case the robot opens the curtains.*

The position in this rule is that privacy is a justifiable trade-off in the face of user distress and resultant psychological harm. The amended impact reflects a positive outcome for autonomy and in preventing harm, but a negative one for privacy. Such trade-offs, we suggest, are negotiated by the stakeholders in determining the most favourable all-things-considered outcome.

To conclude our example, Table 5 considers a few more rules. As indicated by our process, for each rule, we identify the original preliminary rule, the ordering, the defeater(s), any conflicts and impacts, and the SLEEC labels.

5 Conclusion

Rapid progress in the development of autonomous agents has the potential to give rise to applications that can greatly enhance well-being, but an equal potential to do harm. In order to ensure the safe and trustworthy functioning of autonomous agents, it is important to pay careful attention to the social, legal, ethical, and cultural context in which they exist. In order to avoid harm, particularly as these technologies become further integrated in more intimate levels with their users, the imperative to embed SLEEC norms into autonomous agents becomes more pronounced. However, SLEEC norms are often expressed as abstract high-level principles that are not easily reduced to workable rules that an autonomous agent can follow. The five-stage iterative process detailed in this paper describes a method to refine these high-level normative principles into workable rules that must be followed by an autonomous agent that can be trusted to comply with SLEEC norms in a manner that is satisfactory for end users as well as all other stakeholders.

To create autonomous agents that are SLEEC sensitive, we have offered a process to bridge the gap between normative principles and practice. The process can be used to derive SLEEC rules by operationalising normative principles in the guise of evaluative standards, thereby enabling the agent, from a set of possible actions, to select and execute the most normatively favourable action in the intended context, and premised on a range of underlying SLEEC normative principles. Our process only covers one of the main activities of requirements engineering identified by Bennaceur et al. (2019), that is, 'requirements elicitation'. In separate, ongoing project work, we are developing a logic specification language for SLEEC rules, and methods for verifying the compliance of autonomous agents with a formally specified set of SLEEC rules—which Bennaceur et al. (2019) categorise as 'modelling and analysis' and 'assurance' requirements engineering activities, respectively.

Table 5 Examples of stepwise rule elicitation

#	Type	Rule	Impact	Label
0	Prelim	When the user, tells the robot to open the curtains, the robot should open the curtains		
1	Defeater	Unless the user is 'undressed' in which case do not open the curtains and tell the user 'the curtains cannot be opened while the user is undressed'	− A − PH + P + E	Ethical, Social
2	Defeater	Unless the user is 'highly distressed' in which case open the curtains	+ A + PH − P	Ethical, Social
0	Prelim	When dressing the user, close the curtains		
1	Defeater	Unless you are on the 3th floor or above	+ A, + P	Social
0	Prelim	When using emotion recognition system to detect user distress, inform user	+ T + E + PH	Legal, Ethical
0	Prelim	When the cultural dress-preference type is A and gender type is B, dress in clothing item X		
1	Defeater	Unless the user advises otherwise	+ CS + SR + A	Cultural, Social
0	Prelim	Collect only minimum personal information (data minimisation rule)	+ P	Legal

A autonomy, *PH* psychological health (non-maleficence), *P* privacy, *E* explainability, *T* transparency, *CS* cultural sensitivity, *SR* social requirement

In order to reduce high-level principles to workable rules, it is important, as the first stage of the process, to first consider the operating context and design intended for the agent. With this in mind, stakeholders decide on the relevant high-level principles, taking advantage of frameworks such as AI ethics documents, rights-based principles, cultural norms, and appropriate legal codes. At stage 2 these principles, or their proxies, are then mapped to the agent's capabilities and a set of preliminary rules are developed. Stage 3 identifies SLEEC concerns. At this stage the agent determines points of impact within the system-process that may adversely affect a SLEEC norm. Legal concerns and impact assessments are considered. Stage 4 identifies and seeks to resolve conflicts either through assessing trade-offs or generating a compromise. This means that the salience of normative principles within the given context must be taken into account in order to understand which norms and rules take priority within a context. Once that is established, defeaters are generated, where a rule is followed—unless the defeater holds. Defeaters address specific SLEEC concerns with the aim of resolving SLEEC conflicts. Finally, through a process of iteration at stage 5 there is a re-assessment of rules and a generation of complex rules in which rules are amended accordingly and new defeaters advanced. The process of iteration is complete once it is sufficiently reasonable and obvious that the refinement of rules has been exhausted, or there is no further increase in confidence that a new iteration will meaningfully refine the rules further.

Acknowledgements This work was funded by the UKRI project EP/V026747/1 'Trustworthy Autonomous Systems Node in Resilience'. The work of Ana Cavalcanti is funded by the Royal Academy of Engineering, grant CiET1718/45, and the UKRI project EP/V026747/1, UK EPSRC, grants EP/M025756/1 and EP/R025479/1.

Author Contributions Conceptualization: BT, CP, TTA, RC, IH; Methodology: BT, CP, TTA, GN, IH; Formal analysis and investigation: BT, CP; Writing-original draft preparation: BT, CP, TTA, GN, RC, AT; Writing—review and editing: AC, GN, RC, IH, AT; Funding acquisition: AC, RC, IH, AT, Supervision: AC, RC, TTA, AT.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155.

Allen, C., Varner, G., & Zinser, J. (2020). Prolegomena to any future artificial moral agent. In W. Wallach & P. Asaro (Eds.), *Machine ethics and robot ethics* (pp. 53–63). Routledge.

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

Asaro, P. M. (2006). What should we want from a Robot Ethic? *The International Review of Information Ethics, 6,* 9–16.

Audi, R. (2004). *The good in the right: A theory of intuition and intrinsic value*. Princeton University Press.

Bennaceur, A., Tun, T. T., Yu, Y., et al. (2019). Requirements engineering. In S. Cha, R. N. Taylor, & K. C. Kang (Eds.), *Handbook of software engineering* (pp. 51–92). Springer. https://doi.org/10.1007/978-3-030-00262-6_2.

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2018). Social norms. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy, winter* (2018th ed.). Stanford University, Metaphysics Research Lab.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human–Computer Studies, 59*(1–2), 119–155.

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34*(2), 181–186.

BS8611, B. (2016). *Robots and robotic devices, guide to the ethical design and application of robots and robotic systems*. British Standards Institute.

Calinescu, R. (2013). *Emerging techniques for the engineering of self-adaptive high-integrity software* (pp. 297–310). Springer.

Cheng, B. H. C., & Atlee, J. M. (2007). Research directions in requirements engineering. In L. C. Briand & A. L. Wolf (Eds.), *Future of software engineering* (pp. 285–303). IEEE Computer Society. https://doi.org/10.1109/FOSE.2007.17.

Chung, L., Nixon, B. A., Yu, E., et al. (2000). *Non-functional requirements in software engineering*. International series in software engineering (Vol. 5). Springer. https://doi.org/10.1007/978-1-4615-5269-7.

Coşar, S., Fernandez-Carmona, M., Agrigoroaie, R., et al. (2020). Enrichment: Perception and interaction of an Assistive Robot for the Elderly at Home. *International Journal of Social Robotics, 12*(3), 779–805.

Dancy, J. (2004). *Ethics without principles*. Clarendon Press.

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), *Robot law*. Edward Elgar Publishing.

De Voogd, X., Willems, D. L., Onwuteaka-Philipsen, B., et al. (2021). Health care staff's strategies to preserve dignity of migrant patients in the palliative phase and their families: A qualitative study. *Journal of Advanced Nursing, 77*(6), 2819–2830.

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

Dennis, L. A., Fisher, M., & Winfield, A. (2015). Towards verifiably ethical robot behaviour. In: Workshops at the twenty-ninth AAAI conference on artificial intelligence.

Driver, J. (2007). Normative ethics. In F. Jackson & M. Smith (Eds.), *The Oxford handbook of contemporary philosophy*. Oxford University Press.

European Commission. (2019). Ethics guidelines for trustworthy AI. *Publications Office*. https://doi.org/10.2759/177365.

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI Society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2021). An ethical framework for a good AI Society: Opportunities, risks, principles, and recommendations. In L. Floridi (Ed.), *Ethics, governance, and policies in artificial intelligence* (pp. 19–39). Springer.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Foot, P. (2003). Are moral considerations overriding? In P. Foot (Ed.), *Virtues and vices and other essays in moral philosophy*. Oxford University Press.

Formosa, P. (2021). Robot autonomy vs. human autonomy: Social robots, artificial intelligence (AI), and the nature of autonomy. *Minds and Machines, 31,* 1–22.
Future of Life Institute. (2017). ASILOMAR AI principles. Retrieved 31 March, 2022, from https://futureoflife.org/2017/08/11/ai-principles/
Glinz, M. (2007). On non-functional requirements. In: 15th IEEE international requirements engineering conference (pp 21–26). https://doi.org/10.1109/RE.2007.45
Goeldner, M., Herstatt, C., & Tietze, F. (2015). The emergence of care robotics—A patent and publication analysis. *Technological Forecasting and Social Change, 92,* 115–131.
Henderson, D. (2002). Norms, normative principles, and explanation: On not getting is from ought. *Philosophy of the Social Sciences, 32*(3), 329–364.
Holton, R. (2002). Principles and particularisms. *Aristotelian society supplementary volume* (pp. 191–209). Oxford University Press.
Hooker, B. (1996). Ross-style pluralism versus rule-consequentialism. *Mind, 105*(420), 531–552.
Horty, J. F. (2001). *Agency and deontic logic*. Oxford University Press.
Horty, J. F. (2012). *Reasons as defaults*. Oxford University Press.
Hossain, S., & Ishtiaque, Ahmed, S. (2021). Towards a new participatory approach for designing artificial intelligence and data-driven technologies. arXiv:2104.04072
Hurka, T. (2007). Rationality and the good: Critical essays on the ethics and epistemology of Robert Audi. In J. Greco & A. R. Mele (Eds.), *Audi's marriage of Ross and Kant* (pp. 64–72). Oxford University Press.
Jevtić, A., Valle, A. F., Alenyà, G., et al. (2018). Personalized robot assistant for support in dressing. *IEEE Transactions on Cognitive and Developmental Systems, 11*(3), 363–374.
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
Knoks, A. (2020). Defeasibility in epistemology. PhD thesis, University of Maryland, College Park
Latour, B., & Venn, C. (2017). Morality and technology: The end of the means. In M. Jonathan (Ed.), *The ethics of biotechnology* (pp. 87–100). Routledge.
Legros, S., & Cislaghi, B. (2020). Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science, 15*(1), 62–80.
Lindoso, W., Nogueira, SC., Domingues, R., et al. (2021). Visual specification of properties for robotic designs. In: Brazilian symposium on formal methods (pp. 34–52). Springer.
Lutz, C., Schöttler, M., & Hoffmann, C. P. (2019). The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication, 7*(3), 412–434.
Manders-Huits, N., & Van den Hoven, J. (2009). Value-sensitive design.
McDermott, D., & Doyle, J. (1980). Non-monotonic logic I. *Artificial Intelligence, 13*(1–2), 41–72.
McKeever, S., & Ridge, M. (2006). *Principled ethics: Generalism as a regulative ideal*. Oxford University Press.
Menghi, C., Tsigkanos, C., Pelliccione, P., et al. (2019). Specification patterns for robotic missions. *IEEE Transactions on Software Engineering, 47*(10), 2208–2224.
Miyazawa, A., Ribeiro, P., Li, W., et al. (2016). Robochart: A state-machine notation for modelling and verification of mobile and autonomous robots. Tech Rep
Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.
National Institute for Health and Care Excellence. (2013). *Quality standard 50: Mental Wellbeing of Older People in Care Homes*. NICE: Tech. Rep.
Nissenbaum, H. (2014). Respect for context as a benchmark for privacy online: What it is and isn't. Cahier de Prospective 19
OECD.org. (2022). Artificial intelligence. Retrieved 18 March, 2022, from https://www.oecd.org/digital/artificial-intelligence/
Olderbak, S., Sassenrath, C., Keller, J., et al. (2014). An emotion-differentiated perspective on empathy with the emotion specific empathy questionnaire. *Frontiers in Psychology, 5,* 653.
Parfit, D. (1984). *Reasons and persons*. OUP.
Pasquale, F. (2020). *New laws of robotics: Defending human expertise in the age of AI*. Belknap Press.
Pohl, K. (2010). *Requirements engineering—Fundamentals, principles, and techniques*. Springer.
Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science, 11*(4), 481–518.
Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence, 13*(1–2), 81–132.
Reiter, R. (1988). Nonmonotonic reasoning. In H. E. Shrobe (Ed.), *Exploring artificial intelligence* (pp. 439–481). Morgan Kaufmann.

Richardson, H. S. (1990). *Specifying norms as a way to resolve concrete ethical problems* (pp. 279–310). Routledge.
Richardson, H. S. (1997). *Practical reasoning about final ends*. Cambridge University Press.
Ross, D. (2002). *The right and the good*. Oxford University Press.
Scanlon, T. M. (2003). Adjusting rights and balancing values. *Fordham L Rev, 72,* 1477.
Skills for Care. (2013). *Code of conduct for healthcare support workers and adult social care workers in England*. Skills for Care: Tech. Rep.
Smith, P. B., Peterson, M. F., & Schwartz, S. H. (2002). Cultural values, sources of guidance, and their relevance to managerial behavior: A 47-nation study. *Journal of Cross-Cultural Psychology, 33*(2), 188–208.
Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems, 86,* 152–161.
Stratton-Lake, P. (1997). Can Hooker's rule-consequentialist principle justify Ross's prima facie duties? *Mind, 106*(424), 751–758.
The Health Foundation. (2016). *Person-centred Care made Simple: What everyone should know about Person-centred Care*. Health Foundation: Tech. Rep.
Thomas, A. (2000). Consequentialism and the subversion of pluralism. In B. Hooker, E. Mason, & D. E. Miller (Eds.), *Morality, rules, and consequences: A critical reader* (pp. 179–202). Edinburgh University Press.
Thomas, A. (2006). *Value and context: The nature of moral and political knowledge*. Oxford University Press.
Thomas, A. (2011). Another particularism: Reasons, status and defaults. *Ethical Theory and Moral Practice, 14*(2), 151–167.
Tonkens, R. (2012). Out of character: On the creation of virtuous machines. *Ethics and Information Technology, 14*(2), 137–149.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433.
Umbrello, S. (2018). The moral psychology of value sensitive design: The methodological issues of moral intuitions for responsible innovation. *Journal of Responsible Innovation, 5*(2), 186–200.
Umbrello, S. (2019). Beneficial artificial intelligence coordination by means of a value sensitive design approach. *Big Data and Cognitive Computing, 3*(1), 5.
Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics, 1*(3), 283–296.
UNESCO. (2021). Recommendation on the ethics of artificial intelligence. Retrieved 18 March, 2022, from https://unesdoc.unesco.org/ark:/48223/pf0000380455. Document code: SHS/BIO/REC-AIETHICS/2021
Van de Poel, I. (2009). Values in engineering design. In D. M. Gabbay, P. Thagard, J. Woods, & A. W. Meijers (Eds.), *Philosophy of technology and engineering sciences* (pp. 973–1006). Elsevier.
Van de Poel, I., & Kroes, P. (2014). Can technology embody values? In P. Kroes & P. P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 103–124). Springer.
Väyrynen, P. (2009). *A theory of hedged moral principles*. Oxford studies in metaethics.
Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Weinstock, C. B., Goodenough, J. B., & Klein, A. Z. (2013). Measuring assurance case confidence using Baconian probabilities. In: 2013 1st international workshop on assurance cases for software-intensive systems (ASSURE), IEEE (pp. 7–11)
Whittlestone, J., Nyrup, R., Alexandrova, A., et al. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. Nuffield Foundation.
Winfield, A. F., Michael, K., Pitt, J., et al. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]. *Proceedings of the IEEE, 107*(3), 509–517. https://doi.org/10.1109/JPROC.2019.2900622.
Yeung, K. (2020). Recommendation of the Council on Artificial Intelligence (OECD). *International Legal Materials, 59*(1), 27–34. https://doi.org/10.1017/ilm.2020.5.
Yew, G. C. K. (2021). Trust in and ethical design of Carebots: The case for ethics of care. *International Journal of Social Robotics, 13*(4), 629–645.
Zave, P. (1997). Classification of research efforts in requirements engineering. *ACM Computing Surveys (CSUR), 29*(4), 315–321.
Zhang, F., Cully, A., & Demiris, Y. (2019). Probabilistic real-time user posture tracking for personalized robot-assisted dressing. *IEEE Transactions on Robotics, 35*(4), 873–888.

Zytko, D., Wisniewski, P., Guha, S., et al. (2022). *Association for Computing MachineryParticipatory design of AI systems: Opportunities and challenges across diverse users, relationships, and application domains*. https://doi.org/10.1145/3491101.3516506.