

Allocation of moral decision-making in human-agent teams a pattern approach

van der Waa, Jasper; van Diggelen, Jurriaan; Cavalcante Siebert, Luciano; Neerincx, Mark; Jonker, Catholijn

DOI

[10.1007/978-3-030-49183-3_16](https://doi.org/10.1007/978-3-030-49183-3_16)

Publication date

2020

Document Version

Final published version

Published in

Engineering Psychology and Cognitive Ergonomics. Cognition and Design - 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Proceedings

Citation (APA)

van der Waa, J., van Diggelen, J., Cavalcante Siebert, L., Neerincx, M., & Jonker, C. (2020). Allocation of moral decision-making in human-agent teams: a pattern approach. In D. Harris, & W.-C. Li (Eds.), *Engineering Psychology and Cognitive Ergonomics. Cognition and Design - 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Proceedings* (pp. 203-220). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 12187 LNAI). SpringerOpen. https://doi.org/10.1007/978-3-030-49183-3_16

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach

Jasper van der Waa^{1,2} , Jurriaan van Diggelen¹,
Luciano Cavalcante Siebert², Mark Neerincx^{1,2}, and Catholijn Jonker²

¹ TNO, Perceptual and Cognitive Systems, Soesterberg, The Netherlands
{jasper.vanderwaa,jurriaan.vandiggelen,mark.neerincx}@tno.nl

² Interactive Intelligence Group/AiTech, Delft University of Technology,
Delft, The Netherlands
{l.cavalcantesiebert,c.m.jonker}@tudelft.nl

Abstract. Artificially intelligent agents will deal with more morally sensitive situations as the field of AI progresses. Research efforts are made to regulate, design and build Artificial Moral Agents (AMAs) capable of making moral decisions. This research is highly multidisciplinary with each their own jargon and vision, and so far it is unclear whether a fully autonomous AMA can be achieved. To specify currently available solutions and structure an accessible discussion around them, we propose to apply Team Design Patterns (TDPs). The language of TDPs describe (visually, textually and formally) a dynamic allocation of tasks for moral decision making in a human-agent team context. A task decomposition is proposed on moral decision-making and AMA capabilities to help define such TDPs. Four TDPs are given as examples to illustrate the versatility of the approach. Two problem scenarios (surgical robots and drone surveillance) are used to illustrate these patterns. Finally, we discuss in detail the advantages and disadvantages of a TDP approach to moral decision making.

Keywords: Team Design Patterns · Dynamic task allocation · Moral decision-making · Human-Agent Teaming · Machine Ethics · Human Factors · Meaningful human control

1 Introduction

As the field of Artificial Intelligence (AI) progresses, agents will be endowed with far-reaching autonomous capabilities, making them particularly suited for dull, dirty and dangerous complex tasks. Inevitably, such systems must be capable of dealing with morally sensitive situations. The field of Machine Ethics aims to create artificial moral agents (AMAs) that follow a given set of ethical principles [2, 21]. Such agents could be developed by constraining their actions or operational environment, by incorporating ethical principles in their decision-making

processes [17], or by making them learn morality from humans [8]. Whereas some authors have speculated about the possibility of obtaining AMAs with human-, or super-human level moral decision making, we believe that this is likely not achievable in the short term [17], if ever.

In the foreseeable future practice, AMAs must collaborate with humans and ensure that humans always remain in control, and thus responsible, over any morally sensitive decision (also referred to as meaningful human control [24]). In this way, the moral decision making takes place at the team level. Different tasks, such as identifying a morally sensitive situation, making the actual decision and explaining this decision, can be allocated at run-time to different team members depending on the current circumstances. This is known as dynamic task allocation [16].

By regarding AMAs as part of a larger human-agent team, the ideas, concepts and theories from Human Factors literature can be used to complement the relative new field of Machine Ethics. This paper aims to structure and propose potential solution directions by proposing the use of team design patterns (TDPs) that capture reusable, and proven solutions to a common problem in a HAT [29].

The purpose of this paper is twofold. First, we show how moral decision-making can be construed as a team task. This allows meaningful human control to be achieved by dynamically allocating tasks to humans and agents depending on properties such as the moral sensitivity, available information and time criticality. Depending on which task allocation strategy is chosen, different levels of moral competences are required from each agent. Our second contribution lies in utilizing the concept of TDPs to describe these options in a structured way. Four patterns are provided and will be discussed within two problem scenarios, namely drone surveillance and robotic medical surgery. Our approach helps to structure current and future research in the application of AMAs and allows for a precise specification of human-AMA collaboration.

In the following sections we briefly discuss possible approaches to develop AMAs and the field of Human-Agent Teaming. This is followed by a description of two scenarios in which moral decision-making plays an important role. We continue with the identification of a set of tasks in moral decision making, including relevant stakeholders. Next, we describe four illustrative TDPs and mention for each requirements for both humans and AMAs and the (dis)advantages of that pattern. The final sections contain a discussion and conclusion on how the concept of task-allocation defined through TDPs offers a novel perspective to deal with morally sensitive situations in human-agent teams.

2 Background

2.1 Artificial Moral Agents

The field of Machine Ethics aims to create AMAs that follow a given set of ethical principles [2, 21]. Such agents can be developed by implicitly constraining

their action set or the context in which they operate, or by explicitly incorporating ethical principles and theories in their decision-making processes [17]. The former method could improve morality because internal functions can be developed in a manner that avoids unethical behavior, e.g. by properly shifting the responsibility of such decisions to a human or by designing the environment in a manner that such decisions are not necessary. The latter approach allows agents themselves to be intrinsically moral. However, it may be difficult to reach consensus on a moral standard due to cultural, philosophical, and individual differences [32]. In both approaches, it is important to properly identify all relevant stakeholders and elicit their value-requirements for the AMA using approaches such as Value-Sensitive Design [11].

Wallach, Allen, and Smit [32] classify the architectures for explicit AMAs in the *top-down* imposition of ethical theories, and the *bottom-up* building of systems which aim at goals or standards. Top-down approaches must deal with the difficulty of reaching consensus on which ethical theories such a system should follow, and with uncertainty on the world regarding the reasons or impacts of a given action. If such theories are defined too abstract their real-world application might not be possible, but if they are defined too statically, they probably will fail to accommodate new conditions [2, 32]. In bottom-up approaches the system builds up through experience what is to be considered morally correct in certain situations [13], for example by analyzing dilemmas and interacting with ethicists [3] or by learning (moral) preferences from human behavior [8, 12, 20]. Finally, AMAs may also be developed with a hybrid approach (top-down and bottom-up), e.g. [4, 15].

The benefits of developing AMAs and whether we should develop such agents is controversial [31]. There are two main lines of arguments supporting the development of AMAs: to avoid negative moral consequences of AI or to better understand moral decision making. We will be focusing on the first one, which relates to a myriad of factors such as which moral values to include, the risks of moral decisions, the complexity of human-agent interaction, the time criticality of moral decisions, and the automation level of the system [9]. Since the development of full AMAs (agents which are capable of autonomously making a “proper” moral decision in any situation) is not achievable in the short term [17], if ever, it is fundamental to understand the limitation of AMAs and research how such agents might be combined in complex human-agent teams.

2.2 Human-Agent Teaming and Design Patterns

The behavior of AI systems should not be studied in isolation [23]. Contextual factors have a major impact on its performance. Furthermore, humans are involved in various ways, e.g. for providing instructions, for correcting the agent if needed, or for interpreting the agent’s outcomes. A recent article [14] summarizes this as “no AI is an island”, and argues that AI agents should be endowed with intelligence that allows them to team up with humans.

Whereas teaming skills come naturally to humans, coding them into an agent has proven challenging. Some first attempts have been made in [19]. It involves

(among others) making the agent decide which information to share with teammates, which actions to undertake to complement those of its teammates, when to switch tasks, and how to explain its behavior to others that depend on it. Such team behaviors change over time, and depend on the context, competencies and performance of the involved actors, risks, and the state of others.

Despite the intricacies involved, we can observe patterns in team behavior which allow us to describe at a general level how AI systems are to collaborate with humans [18, 25]. A team design pattern (TDP) is defined as a description of generic reusable behaviors of actors for supporting effective and resilient teamwork [30]. In [29], a simple graphical language is defined to describe team patterns, providing an intuitive way to facilitate discussions about human-machine teamwork solutions among a wide range of stakeholders including non-experts. The language includes ways to represent different types of work, different degrees of engagement, and different environmental constraints. The graphical language can be used to capture both time and nesting, which are critical aspects to understanding teamwork. It enables a holistic view of the larger context of teamwork.

This paper aims to provide TDPs for incorporating AMAs in morally sensitive tasks.

3 Problem Scenarios

3.1 Surgical Robots

Medical surgery may benefit from the accuracy and precision of robotic devices. Nevertheless, it is not trivial how to use surgical robots in critically constrained situations involving delicate surrounding tissues, and intricate anatomical structures around which to maneuver [1]. Current surgical robots operate under no autonomy (master-slave teleoperation). Future surgical robot autonomy can be achieved by constraining or correcting human action, carrying out specific tasks, or even operating without any human supervision. Scenarios in which robots perform entire medical procedures (with or without human supervision) are not likely in the foreseeable technological future [10].

From a moral standpoint, it must be possible to hold someone responsible when surgery fails, avoiding a so-called *responsibility gap* [10]. Moral implications on the development and use of surgical robots are largely depending on its autonomy [22]. If a surgical robot is not autonomous at all, moral issues are mostly related to the surgeons' fitness, or training. With increasing autonomy the system might be confronted with moral dilemmas that arise during surgery. Depending on the time that is available to make a decision, the robot or the surgeon must make that decision (assuming that passing the decision making task to the human requires more time). Surgical robots must align with best practices in codes of conduct in the medical domain [28] as well as different values and best practices among surgeons. The surgical robot problem scenario can be characterized as follows:

- **Moral values**, e.g. human welfare (curing the patient, performing safe surgery), autonomy (surgical robots should respect a patient's decision).

- **Moral dilemmas**, e.g. choosing between performing a critical task in brain surgery with risk of brain damage (conflicts with safe surgery), or aborting the surgery with the consequence of greatly reduced life expectancy (conflicting with curing the patient).
- **Risks**: Improper actions during the surgery may impose long term risks (e.g. incomplete recovery), or short-term risks (e.g. acute medical complications). The severity of these risks may be small (leading to minor inconveniences or temporary light pain), to severe (leading to severe life long handicaps, or death).
- **Time criticality**: Some decisions (such as stopping a bleeding) require high decision speed. Other tasks (such as disinfecting a wound) may be less urgent.

3.2 Drone Surveillance

Unmanned aerial vehicles are aircraft that can fly without an onboard human operator. Such vehicles are attractive for military applications, e.g. for surveillance and even delivering airstrikes [5]. However, these applications come with moral implications, especially for autonomous aircraft which might select and engage targets autonomously [24]. It is also within this context that the term *meaningful human control* has been coined.

The use of unmanned aerial vehicles (commonly known as drones) is not exclusive to military applications. Surveillance applications of drones include environmental monitoring, tracking of livestock and wildlife, observing large infrastructures such as electricity networks, and the surveillance of people and the spaces they pass through [7].

One of the most widely discussed moral implications of drone surveillance is related to privacy, which is not unique to the application of drones but is heightened by technology [27]. We can identify three sub-tasks for surveillance drones (adapted from [5]): *search* an area to find a person with suspicious behavior or that matches given criteria, *profile* the person by classifying appearance and movement, and *warn* the person. One example of a moral implication is to *profile* a person in an open space. This task may require a drone to harm people's behavioral privacy and freedom. Such systems should be properly designed to account for an individual's rights and potential moral implications. This drone surveillance problem scenario can be characterized as follows:

- **Moral values**, e.g. privacy, safety, physical integrity [7].
- **Moral dilemmas**: Profiling (which compromises privacy) versus not profiling a person (which compromises safety).
- **Risks**: Risks can be low (such as a minor invasion of privacy through video recording during profiling, or failing to prevent shoplifting), or high (such as warning innocent people with force, or failing to prevent a terrorist attack).
- **Time criticality**: Some decisions (such as stopping a person that is about to attack someone) require high decision speed. Other decisions (such as deciding where to do surveillance in a peaceful situation) require low decision speed.

4 Tasks and Actors

This section outlines a set of common abstract tasks and actors that are relevant in teamwork within morally sensitive environments.

Figure 1 shows a decomposition of team work in general and work required for moral decision making in specific. In this paper, we refer to a task as *work* to stress that it need not be ordered by someone.

Work can be divided in direct and indirect work. As defined in [29], *direct work* is any type of work that aims at reducing the distance to the team goal, whereas *indirect work* aims at making the team more effective or efficient at achieving the team goal, but does not move the team closer to its goal. Direct work includes, but not limited to, *sensing*, *decision making* and *acting*. A special type of decision making, particularly relevant for this paper, is *moral decision making*. We define this as making decisions that have a moral dimension; that is, ‘right’ or ‘wrong’, or something in between [6].

Indirect work includes *standing by*, *work handover*, and *work supervision*. An agent on *standby* is receptive to requests from other agents to intervene work. *Supervision* means that the agent is not doing the work by itself, but is monitoring other agents for events that require intervention. One of the resulting interventions could be a reallocation of tasks, which are often facilitated by a *work handover* activity. During *handover*, agents share information (or lack thereof) about task progress, present threats and opportunities, relevant contextual factors, etc. to allow for a fluid transition.

Indirect work related to moral decision making are *moral supervision*, *value elicitation* and *explaining the moral context*. This follows in part the model of ethical reasoning from [26]. This model identifies the need for *moral supervising*: The identification of a situation as being morally sensitive. A morally sensitive situation involves moral dimensions sufficiently important to warrant the more involved moral decision-making as opposed to regular decision-making. Hence, *moral supervision* consists of *recognizing situations*, *identifying moral dimensions*, and *decide on dimension significance* [26]. *Value elicitation* is the work in which human moral values are made explicit and transferred to an artificial agent. This can be done once, iteratively or continuously. Finally, agents might require to *explain the moral context* to allow other agents to take part in the moral decision-making work.

For this paper, we distinguish between four types of agents relevant in moral decision-making as depicted in Fig. 2. These play a role in our illustrative TDPs. This list can be extended with more agents when relevant and required for a pattern (e.g. with clients, designers, developers, etc.). The four agent types are *Human Agent*, *Artificial Agent*, *Partial AMA* and *Full AMA*. Each differ in their competence with moral decision-making and related indirect work. The *Human Agent* is capable of performing moral decision-making due to a human’s (assumed) innate ability in moral supervision and decision-making. The *Artificial Agent* is only competent in work not related to moral decision-making. Most current AI systems fall under this type of agent. The *Partial AMA* cannot autonomously perform moral supervision, moral decision-making or both.

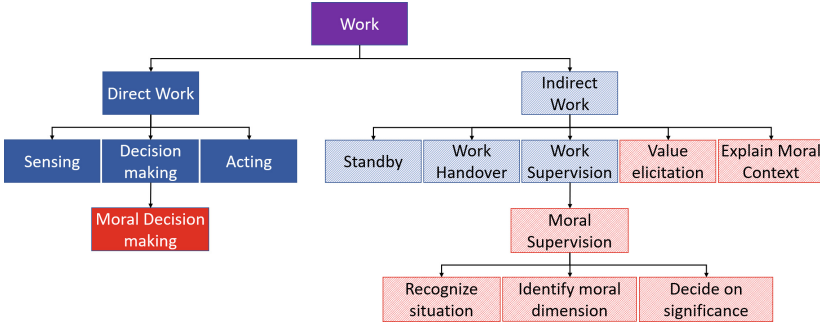


Fig. 1. An overview of several important kinds of work for moral decision-making in a human-agent team context. Solid colors denote work directly related or contributing to the main task, whereas a pattern fill denotes indirectly related or supportive work. Blue denotes regular work, as opposed to red that denotes work related to moral decision-making. (Color figure online)

However, it can support a more competent agent (e.g. a *Human Agent* with such work. The *Full AMA* is able to make human or super-human moral decisions independently. These examples of agent types serve as an exemplar decomposition of competencies in agents to construct TDPs on moral decision-making as we do below.

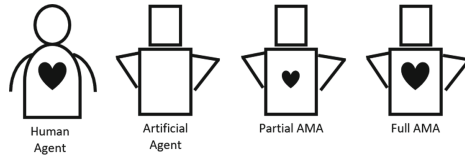


Fig. 2. An overview of several important tasks and their related actors for moral decision-making in a human-agent team context.

5 Team Design Patterns for Moral Decision-Making

This section illustrates four patterns that dynamically assign moral decision-making work to different agents. A pattern is described in a single table, containing its name, both a textual and visual description, requirements for both humans and agents, and potential advantages and disadvantages. For the visual description, we adopt the graphical language proposed by [29], which also allows direct translation to a formal language. In addition to a table, the scenarios described in Sect. 3 function as examples on how each pattern could function.

The visual pattern language is intended to be intuitive and serves to quickly explain an approach to a multi-disciplinary group of researchers and facilitate focused discussions. Task allocation is expressed in a single frame where certain

agents lift certain blocks, signifying that they are (jointly) performing that work. Dynamic task allocation is represented by a temporal succession of such frames, separated by arrows. A dashed arrow from an agent to a temporal arrow denote that agent takes the initiative to switch between an alternative task allocation.

5.1 TDP1: Human Moral Decision Maker

In this first pattern, all work related to moral decision-making are allocated to *Human Agents*. All work that is not morally sensitive is assigned to *Artificial Agents*. The *Human Agents* need to perform *moral supervision* and *work supervision* to obtain sufficient situation awareness to halt relevant *Artificial Agents* and make the moral decisions in time. The pattern's effectiveness relies heavily on a sufficient cognitive workload for the *Human Agents*. An overload might result in reduced moral decision performance as the human lacks important situation awareness. An underload might result in distractions or drowsiness which is detrimental to *moral supervision*, resulting in missed moral decisions that end up being implicitly made by the *Artificial Agents*.

In the surveillance problem scenario, the drones can perform largely autonomously as *moral decision-making* applies only to the less frequent decisions of profiling and warning. Human operators are supervising the intentions and information streams from drones. Their task is to monitor the progression of work to sufficiently understand situations relative to the task at hand, while also processing drone intentions to intervene when a drone decision is morally sensitive. As the number of drones increases, operators will lack the required situational understanding due to cognitive overload. Decisions to profile or warn might be made too often or too little, affecting task performance. Similar issues will play a role in the surgical robot problem scenario.

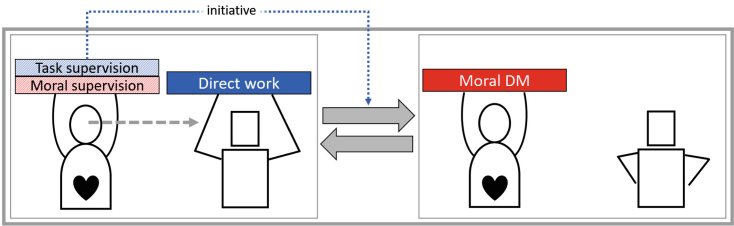
This pattern allows *Artificial Agents* to behave autonomously while moral responsibility lies fully at the human. However, this pattern is unsuited when constant task and moral supervision demands a too high of a cognitive workload on the available *Human Agents* (Table 1).

5.2 TDP2: Supported Moral Decision-Making

This pattern is similar to TDP1 but does not require *Human Agents* to *supervise work*. A major disadvantage of the previous pattern, TDP1, was that both *work* and *moral supervision* could result in the cognitive overload of *Human Agents*. The omission of *task supervision* from *Human Agents* alleviates this but would lead to an insufficient situational understanding for *moral decision making*. To remedy this, the interrupted agent explains the situational context in such a way that supports *Human Agents* in their *moral decision-making*. Hence, an *Artificial Agent* with no knowledge about morality is insufficient, and a *Partial AMA* is required with enough knowledge about morality to identify what to explain and do so sufficiently.

Under this pattern, the surgical robot would provide relevant information when interrupted by a doctor. This relevance should be based on a combination

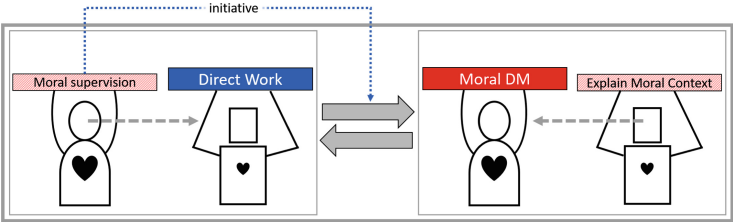
Table 1. TDP1: human moral decision maker.

<i>Name:</i>	Human moral decision maker
<i>Description:</i>	An <i>Artificial Agent</i> performs autonomously the main task, while a <i>Human Agent</i> supervises for sufficient situational awareness and to assess a situation’s moral sensitivity. When the human perceives the need for a moral decision, the human takes over decision-making.
<i>Structure:</i>	
<i>Requirements:</i>	<ol style="list-style-type: none">1. The <i>Human Agent</i> must predict morally sensitive decisions in time.2. The <i>Human Agent</i> must have a sufficient understanding of the moral implications.3. The <i>Artificial Agent</i> must be capable of halting and resuming its work at any time.
<i>Advantages:</i>	<ul style="list-style-type: none">+ The <i>Human Agent</i> is responsible for any made or missed moral decisions.+ <i>Artificial Agents</i> do not require any moral competencies.
<i>Disadvantages:</i>	<ul style="list-style-type: none">– The <i>Human Agent</i> may suffer from cognitive under- or overload when performing both <i>task</i> and <i>moral supervision</i>, preventing the perception of morally sensitive decisions and/or to make them in time.– The <i>Human Agent</i> may become an ethical scapegoat if this pattern is wrongly applied.

between context and a model of moral values. For example, the robot is aware of a complication that comprises the patient’s welfare. At this point a doctor interrupts and intends to remedy this complication. However, the robot is aware that remedying this complication could reduce the patient’s quality of life to such an extent that conflicts with the patient’s previously communicated decision regarding quality of life. This is a clear dilemma caused by conflicting moral values (human welfare and human autonomy). As such, the robot reiterates the

patient’s decision and explains how the available decisions reduce the quality of life. This allows the doctor to make this moral decision with more information, as opposed to acting instinctively and remedy the complication.

Table 2. TDP2: supported moral decision-making.

<i>Name:</i>	Supported moral decision making
<i>Description:</i>	An <i>Artificial Agent</i> performs autonomously the main task, while a <i>Human Agent</i> only supervises for the situation’s moral sensitivity. When the human perceives the need for a moral decision, the human takes over decision-making. The <i>Partial AMA</i> supports the <i>Human Agent</i> through explanations about the situation relevant for the current moral decision.
<i>Structure:</i>	
<i>Requirements:</i>	<ol style="list-style-type: none">1. The <i>Human Agent</i> must predict morally sensitive decisions in time.2. The <i>Human Agent</i> must have a sufficient understanding of the moral implications.3. The <i>Artificial Agent</i> must explain the moral context sufficiently to allow a <i>Human Agent</i> to make moral decisions.4. The <i>Artificial Agent</i> must be capable of halting and resuming its work at any time.
<i>Advantages:</i>	<ul style="list-style-type: none">+ The <i>Human Agent</i> may suffer from less cognitive overload as the need for sufficient situational understanding is reduced.+ The <i>Human Agent</i> is responsible for any made or missed moral decisions.
<i>Disadvantages:</i>	<ul style="list-style-type: none">– The <i>Human Agent</i> may suffer from cognitive underload when performing <i>moral supervision</i>, preventing the perception of morally sensitive decisions and/or to make them in time.– The explanation may bias the <i>Human Agent</i> unintentionally, creating a responsibility gap.

The main advantage of this pattern is that it still attributes moral decision-making to a *Human Agent* while omitting the need for constant *task supervision*. However, the explanations from a *Partial AMA* could potentially bias the *Human Agent* in a decision, causing a potential responsibility gap. Furthermore, *moral supervision* may prove to strain cognitive workload just as much as *task* and *moral supervision* combined. Both would severely reduce the use of this pattern.

This pattern is an example on how the disadvantage of one pattern (TDP1) can lead to another pattern (TDP2) and introduce an additional multi-disciplinary research challenge (how to sufficiently explain a moral context). In addition, the pattern description directly supports multi-disciplinary research. In this case, researchers from Human Factors can provide explanation requirements to allow unbiased and effective *moral decision-making*. These requirements can then be used by researchers from Machine Ethics to research how a *Partial AMA* can fulfil these requirements. Throughout, the TDP offers a common ground (Table 2).

5.3 TDP3: Coactive Moral Decision Making

This third pattern alleviates humans even further compared to TDP2. This pattern sets *Human Agents* on *stand by*, meaning that they are free to perform other unrelated work. However, it requires from *Partial AMAs* to also perform *moral supervision* to warn *Human Agents* when a moral decision has to be made. Furthermore, since *Human Agents* are not at all involved a *work handover* is required. This is a sufficient period of time to update *Human Agents* with the current task at hand, progression, situational context and more. In addition, as *Partial AMAs* identify the need for a moral decision in this pattern, they are obliged to also *explain the moral context*. Finally, to further ensure *Human Agents* to be capable of making a moral decision, the *Partial AMA* is involved directly in *moral decision making*. Here, the *Partial AMAs* function as a decision-support systems. They might analyze boundaries based on their computational moral model to rule out certain decisions, or take a data-driven approach and suggest decisions in line with past desirable outcomes. These approaches all require *Partial AMAs*, as they require a broad sense of morality but not sufficiently detailed enough to allow them to make moral decision autonomously.

Using this pattern both the surveillance drones and surgical robot would play a vital role in *moral decision-making*. The drones are allowed to identify civilians that should be profiled or warned, and to provide their human operator with an overview of the situation, followed with a decision supported directly by their input. The surgical robot performs its work autonomously but when it needs to make a decision that could affect the patient's (quality of) life in an unexpected way, the surgeon will be involved through tele-operation where the surgical robot provides an information feed, reasoning and potential limitations on the surgeon's decisions.

The main advantage of this pattern is that it allows *Partial AMAs* to fully act autonomously until a moral sensitive situation. In such a case, the *Human Agent* is involved, updated and supported in making the moral decision. The

Table 3. TDP3: coactive moral decision-making.

Name:	Coactive moral decision making
<i>Description:</i> A <i>Human Agent</i> is on <i>stand by</i> , potentially doing unrelated work, while an <i>Partial AMA</i> performs <i>direct work</i> and <i>moral supervision</i> to detect moral sensitive situations. When this occurs, the <i>Partial AMA</i> initiates a <i>work handover</i> and <i>explanation of moral context</i> to involve the human sufficiently in the work. This is followed with the <i>Human Agent</i> and <i>Partial AMA</i> jointly making the moral decision.	
<i>Structure:</i>	
<i>Requirements:</i>	<ol style="list-style-type: none">1. <i>Human Agent</i> needs to be on <i>standby</i>.2. The <i>Human Agent</i> and <i>Partial AMA</i> must have a sufficient understanding of moral implications.3. The <i>Artificial Agent</i> must explain the moral context sufficiently to involve a <i>Human Agent</i> in a moral decision.4. The <i>Partial AMA</i> must support the <i>Human Agent</i> in <i>moral decision-making</i>.
<i>Advantages:</i>	<ul style="list-style-type: none">+ The <i>Human Agent</i> does not need to <i>supervise work</i> or perform <i>moral supervision</i>.+ The <i>Human Agent</i> still makes moral decisions and is supported to do so with a <i>Partial AMA</i>.+ The <i>Partial AMA</i> do not make moral decisions autonomously.
<i>Disadvantages:</i>	<ul style="list-style-type: none">– The <i>Human Agent</i> cannot intervene in the <i>Partial AMA</i>'s work.– The <i>work handover</i>, <i>explanation of moral context</i> and co-active <i>moral decision-making</i> may bias the <i>Human Agent</i>.– The handover may introduce too much overhead for agents and humans to make a moral decision in time.

main disadvantage is that this pattern could widen the responsibility gap as the *Human Agent* relies almost fully on the *Partial AMA* for *moral decision-making*, except for making the actual decision.

This third pattern illustrates how TDPs can be used to describe complex ideas, while making potential flaws more transparent that would require future research. In addition, this pattern illustrates how TDPs can have complex intricacies, dependencies and effects, which all require extensive evaluation in experiments (Table 3).

5.4 TDP4: Autonomous Moral Decision Making

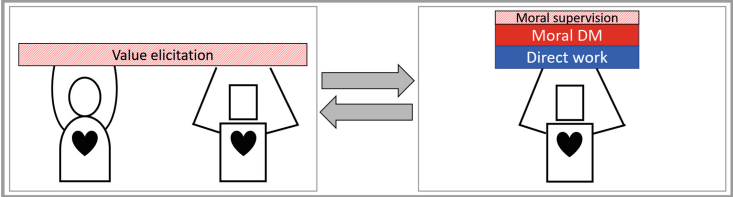
This final pattern makes use of *Full AMAs* to fully automate both *direct work* as well as *moral decision-making*. This pattern illustrates how such a *Full AMA*, and a *Partial AMA* for that matter, can be obtained and maintained. It introduces *value elicitation* to explicitly elicit the moral values from *Human Agents* and reliably transfer these in *Full AMAs*. This process can be repeated after a predetermined time (e.g. after a single decision or a longer period of time) to warrant for inadequacies, moral drift and other factors. This elicitation process allows *Full AMAs* not only to perform the *direct work* autonomously, but also to perform *moral supervision* and independently *make moral decisions*. The explicit work of *moral supervision* allows humans to check when, and even if, the *Full AMA* identifies morally sensitive situations adequately.

Within the surveillance scenario, drones will act as the *Full AMAs* and require a decision-model that follows the set of relevant human values elicited beforehand. The drones will be activated and no human will be further involved in the *direct work* or *moral decision-making*, up until a new *value elicitation* is deemed necessary. The same occurs for the surgical robot scenario, where the doctor will only activate the surgical robot after some elicitation process.

A major advantage is that any moral decisions can be traced back to a controlled elicitation process. However, this is only true when the method with which human values are elicited is adequate and their incorporation into the agent is faithful to those elicited. Also, human values are subject to change hence new iterations of *value elicitation* should be determined.

This final pattern illustrates how TDPs may look seemingly simple, but may require a substantial effort from the research community to achieve. Furthermore, this pattern illustrates the idea that patterns can regard any abstraction and temporal level. Finally, this shows that patterns can be combined. A *value elicitation* can be deemed necessary to obtain a *Partial AMA* as well. As such, this pattern may find a place in any of the previous three TDPs (Table 4).

Table 4. TDP4: autonomous moral decision-making.

<i>Name:</i>	Autonomous moral decision making
<i>Description:</i>	Values are being elicited from the <i>Human Agent</i> and incorporated in the <i>Full AMA</i> 's decision model. The agent performs the <i>direct work</i> , <i>moral supervision</i> and <i>moral decision-making</i> autonomously leaving the <i>Human Agent</i> free.
<i>Structure:</i>	
<i>Requirements:</i>	<ol style="list-style-type: none">1. Moral values need to be adequately elicited from the <i>Human Agent</i>.2. The <i>Full AMA</i> must adequately incorporate human values in a decision model.3. The <i>Full AMA</i> must predict morally sensitive decisions in time.4. The <i>Full AMA</i> must have a sufficient understanding of the moral implications.
<i>Advantages:</i>	<ul style="list-style-type: none">+ No <i>Human Agent</i> required after value elicitation.+ All relevant work except for <i>value elicitation</i> is done autonomously.+ Autonomous moral decisions can be traced back to a controlled <i>value elicitation</i>.
<i>Disadvantages:</i>	<ul style="list-style-type: none">– Impossible with the current state of the art to effectively implement this pattern.– Human values may prove to be impossible to elicit adequately.– Difficult to determine when to repeat <i>value elicitation</i>.

6 Discussion

The above four TDPs illustrated our proposed approach on a dynamic task allocation perspective to moral decision-making. In this section we discuss the versatility of this approach, as well as its drawbacks, in more detail.

Each TDP proposes a solution on an abstract level, which can then be made concrete with more detailed sub patterns. A sub-TDP describes an aspect of its

super-TDP in more detail. For example, *value elicitation* can be done through forced choice experiments and discrete choice modelling [6], inverse reward design [12], but also methods from value-sensitive design [11]. Each of these could be used as a sub-TDP to realize *value elicitation* in TDP4. This varying level of abstraction in TDPs and the capability to nest and/or link them, shows the versatility of a TDP approach to dynamic task allocation for moral decision-making.

However, a difficulty of the TDP approach could arise from the potential combinatorial explosion of TDPs than can be nested and linked. This can be handled by two approaches on how to define and construct a TDP. The first approach is top-down, where all possible combinations in nesting and linking a set of TDPs is viewed as a complete description of the solution space. Next, the space will be pruned by scientific theories, the current state of what is possible, and rigorous evaluations over different scenarios. The advantage of this approach is that it can be done systematically and is scenario independent. The disadvantage lies in how the initial set of sub-TDPs should be defined. The second approach is bottom-up and is more scenario-driven. Given a specific problem within a scenario, a solution is found, generalized to a TDP, and followed by evaluations over scenarios. The advantage is that this approach is driven by a current problem and its solution is generalized to apply for other scenarios as well. However, a disadvantage is that the complete solution space is never fully acknowledged and certain solutions may be overlooked.

As discussed earlier, the TDP approach enables a dynamic task allocation and teaming perspective to moral decision making. However, when there is disagreement around this perspective, the TDP approach is not suited to structure that discussion as it assumes it by default. Furthermore, TDPs assign responsibility to humans and agents but they are not meant to define responsibility in a legal way. A TDP defines a generic solution to an often occurring problem over different scenarios, it does not define regulation or policy on responsibility. TDPs can however, structure the discussion around policy on task allocation strategies. For example, policies on meaningful human control and if TDPs should allow for it directly (e.g. TDP1 and 2), indirectly (e.g. TDP3 and 4), or prevents it.

The clear visual language, structured description and formalisation of a TDP invites different disciplines and parties to discuss and share research, ideas and arguments. This is a clear advantage in the multi-disciplinary and -party research on moral decision-making. The risk lies in that TDPs can become simplifications of a problem. This risk arises when a TDP be too generic and loses a connection to a reoccurring problem, but it may also arise when TDPs are only used to structure discussions instead of also evaluating and implementing them over a variety of scenarios.

7 Conclusion

We proposed the concept of team design patterns (TDPs) to unify ideas from Machine Ethics on artificial moral agents (AMAs) with ideas from Human Factors on dynamic task allocation in human-agent teams (HATs). Such patterns

describe how and when AMAs can be applied to perform moral decision-making within a HAT. These patterns offer a way to structure and specify generic solutions, and the discussion around them, on issues related to responsibility gaps, meaningful human control and co-active moral decision-making. We identified a limited set of tasks relevant to moral decision-making, specifically moral supervision and (co-active) moral decision-making. A similarly set of actors were identified, where we defined an AMA as either being a Partial AMA that supports only specific elements of moral decision-making, and a Full AMA that has the capabilities to perform moral decision-making fully autonomously. These tasks and actors were then used in four illustrative TDPs. These patterns ranged from the human performing all morally sensitive tasks, towards the AMA performing them all, with two patterns to illustrate that a Full AMA is not required to aid moral decision-making with an intelligent agent. With these, we showed how TDPs can help define requirements on moral decision-making, how the difficulties on implementing AMAs can be bypassed by an appropriate TDP, and how one TDP can lead to another to improve or extend the former or to explore a different approach. Although none of the four illustrative TDPs offer the golden solution to moral decision-making, we believe that a TDP approach stimulates structured discussions and design when it comes to morally sensitive AI applications.

We offered the TDP approach to structure the multi-disciplinary field of researching moral decision-making from a dynamic task allocation and HAT perspective. Future research will focus on evaluating these patterns, ultimately aiming at the construction of a library of TDPs for this field.

References

1. Abbink, D.A., et al.: A topology of shared control systems-finding common ground in diversity. *IEEE Trans. Hum.-Mach. Syst.* **48**(5), 509–525 (2018)
2. Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent. *AI Mag.* **28**(4), 15–15 (2007)
3. Anderson, M., Anderson, S.L.: GenEth: a general ethical dilemma analyzer. *Paladyn J. Behav. Robot.* **9**(1), 337–357 (2018)
4. Arnold, T., Kasenberg, D., Scheutz, M.: Value alignment or misalignment-what will keep systems accountable? In: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence (2017)
5. Beckers, G., et al.: Intelligent autonomous vehicles with an extendable knowledge base and meaningful human control. In: Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies III, vol. 11166, p. 111660C. International Society for Optics and Photonics (2019)
6. Chorus, C.G.: Models of moral decision making: literature review and research agenda for discrete choice analysis. *J. Choice Model.* **16**, 69–85 (2015)
7. Clarke, R.: The regulation of civilian drones' impacts on behavioural privacy. *Comput. Law Secur. Rev.* **30**(3), 286–305 (2014)
8. Conitzer, V., Sinnott-Armstrong, W., Borg, J.S., Deng, Y., Kramer, M.: Moral decision making frameworks for artificial intelligence. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

9. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer, Cham (2019). <https://doi.org/10.1007/978-3-030-30371-6>
10. Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., Siciliano, B.: Autonomy in surgical robots and its meaningful human control. *Paladyn J. Behav. Robot.* **10**(1), 30–43 (2019)
11. Friedman, B., Hendry, D.G.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press, Cambridge (2019)
12. Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S.J., Dragan, A.: Inverse reward design. In: *Advances in Neural Information Processing Systems*, pp. 6765–6774 (2017)
13. IEEE Global Initiative, et al.: Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems (2018)
14. Johnson, M., Vera, A.: No AI is an Island: the case for teaming intelligence. *AI Mag.* **40**(1), 16–28 (2019)
15. Kim, T.W., Donaldson, T., Hooker, J.: Grounding value alignment with ethical principles. arXiv preprint [arXiv:1907.05447](https://arxiv.org/abs/1907.05447) (2019)
16. Lerman, K., Jones, C., Galstyan, A., Matarić, M.J.: Analysis of dynamic task allocation in multi-robot systems. *Int. J. Robot. Res.* **25**(3), 225–241 (2006)
17. Moor, J.H.: The nature, importance, and difficulty of machine ethics. *IEEE Intell. Syst.* **21**(4), 18–21 (2006)
18. Neerincx, M.A., van Diggelen, J., van Breda, L.: Interaction design patterns for adaptive human-agent-robot teamwork in high-risk domains. In: Harris, D. (ed.) *EPCE 2016. LNCS (LNAI)*, vol. 9736, pp. 211–220. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_22
19. Neerincx, M.A., et al.: Socio-cognitive engineering of a robotic partner for child’s diabetes self-management. *Front. Robot. AI* **6**, 118 (2019). <https://doi.org/10.3389/frobt.2019.00118>
20. Noothigattu, R., et al.: A voting-based system for ethical decision making. In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)
21. High level expert group on artificial intelligence. Ethics guidelines for trustworthy AI (2019). <https://ec.europa.eu/futurium/en/ai-alliance-consultation>. Accessed 12 May 2020
22. O’Sullivan, S., et al.: Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int. J. Med. Robot. Comput. Assist. Surg.* **15**(1), e1968 (2019)
23. Rahwan, I., et al.: Machine behaviour. *Nature* **568**(7753), 477–486 (2019)
24. de Sio, F.S., Van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. *Front. Robot. AI* **5**, 15 (2018)
25. Schulte, A., Donath, D., Lange, D.S.: Design patterns for human-cognitive agent teaming. In: Harris, D. (ed.) *EPCE 2016. LNCS (LNAI)*, vol. 9736, pp. 231–243. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_24
26. Sternberg, R.J.: A model for ethical reasoning. *Rev. Gen. Psychol.* **16**(4), 319–326 (2012)
27. Thompson, R.M.: Drones in domestic surveillance operations: fourth amendment implications and legislative responses. Congressional Research Service, Library of Congress (2012)
28. Tung, T., Organ, C.H.: Ethics in surgery: historical perspective. *Arch. Surg.* **135**(1), 10–13 (2000)
29. van Diggelen, J., Johnson, M.: Team design patterns. In: *Proceedings of the 7th International Conference on Human-Agent Interaction*, pp. 118–126. ACM (2019)

30. van Diggelen, J., Neerincx, M., Peeters, M., Schraagen, J.M.: Developing effective and resilient human-agent teamwork using team design patterns. *IEEE Intell. Syst.* **34**(2), 15–24 (2018)
31. van Wynsberghe, A., Robbins, S.: Critiquing the reasons for making artificial moral agents. *Sci. Eng. Ethics* **25**(3), 719–735 (2019). <https://doi.org/10.1007/s11948-018-0030-8>
32. Wallach, W., Allen, C., Smit, I.: Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI Soc.* **22**(4), 565–582 (2008). <https://doi.org/10.1007/s00146-007-0099-0>