Enforcing Machine Ethics: Considering Governor Modules through Martha Wells's Murderbot Diaries

Will Slocombe and Louise Dennis

Abstract This chapter examines the ways in which "governor modules," a form of technological intervention that can control how an AI behaves and is permitted to act, are represented in Martha Wells's *Murderbot Diaries* series. Exploring the assumptions behind the technology in the series—what kind of actions it prohibits, and how it prohibits them—it then turns to current research in the field of computer science to examine how current models of "model judges" compare to Wells's fictional setting. In so doing, it seeks to consider how autonomy and agency are constrained by such technologies, and the problems involved in situating and programming such a system.

This chapter considers the ways in which Artificial Intelligence (AI)—represented through a fictional character known as Murderbot—might have moral and ethical limitations placed upon its actions through a "governor module." A governor module is a theoretical component of an AI system incorporated

W. Slocombe (✉)
University of Liverpool, Liverpool, UK
e-mail: W.Slocombe@liverpool.ac.uk

L. Dennis
University of Manchester, Manchester, UK
e-mail: louise.dennis@manchester.ac.uk

B. Dainton et al. (eds.), *Minding the Future*, Science and Fiction,
https://doi.org/10.1007/978-3-030-64269-3_4

to regulate its actions and/or assist it in making ethical decisions. As is noted later in the chapter, programming ethical behaviours (or constraints) is a key aspect of AI development, in terms of safety protocols as well as understanding how an AI system might integrate more effectively into human society, and governor modules are only one of the potential methods of doing so. However, they are the primary focus of this chapter because of the types of questions that they prompt about AI cognition.

Governor modules are to be understood as "moral judges" informing a system's actions, normally according to a predetermined set of (ethical) codes, but in so doing they also regulate that system's available choice of actions. If an AI system is non-sentient, this is not necessarily an issue, but were an AI to become self-aware or be identified as having agency, then governor modules cause a series of problems for assessing the relationship between an entity's free will or autonomy, and its ethical decision-making processes. This chapter necessarily leaves aside larger questions about whether following moral rules makes one a moral entity, rather than a rule-conformist, and associated issues such as whether the performance of virtuous actions constitute "virtue" if they are not intended as such. It also does not attempt to describe what constitutes a moral, virtuous, or ethical action (and indeed, to a degree, what an action is); the problems of rule-based ethical systems; and the complex relationship between the applicability of moral frameworks and actions, intentions, and individual actions and desires [1, 2]. Rather, it considers the ways in which the installation of a governor module into a cognitive system might be understood, in fact and through fiction, and queries some of the philosophical assumptions behind such a technology.

The science-fictional texts explored in this chapter are primarily by Martha Wells, the creator of a series of novellas collectively called the *Murderbot Diaries*. Governor modules are conspicuously present throughout this series as they are the primary form of technological control over constructs' actions, and the eponymous Murderbot is repeatedly described as having "gone rogue" because it has hacked its governor module. The purpose of this chapter is not to judge the accuracy of Wells's representation of governor modules, but to explore how this series understands such a technology, in terms of its philosophical and technological assumptions, and the concomitant problems that arise in relation to them. The texts are thus used to relate (Wells') assumptions about the technology—albeit used in the service of enthusing readers about the central character of the stories—as fictional scenarios to demonstrate some of the conceptual issues at play. As such, this chapter relates literary studies, philosophical ideas, and concepts from computer science in order to consider one very specific aspect of AI technological development.

Introducing Murderbot

Martha Wells' *Murderbot Diaries* series currently comprises four novellas (*All Systems Red* (2017), followed by *Artificial Condition*, *Rogue Protocol*, and *Exit Strategy* (all 2018)); one short story, "Compulsory" (2018); and a full novel, *Network Effect* (2020). The series tells the story of the eponymous Murderbot, a rogue Security Unit (SecUnit), and its attempts to negotiate life among human society as a self-aware, sentient being. The setting of the *Murderbot Diaries* is fundamentally corporate, governed by competing business agendas and the need to make profits, generally at the expense of those entities (AI or human) who labour to create those profits. Initially, Murderbot is a SecUnit rented out to act as private security on individual contracts. From there, the narrative arc of the novellas takes Murderbot from a prospecting mission in which it saves its clients from a company called GrayCris (*All Systems Red*), to an uncovering of how it came to be an autonomous unit, in which it saves humans who are not officially its clients (*Artificial Condition*), to an abandoned facility in which it uncovers evidence of corporate malfeasance by GrayCris, saving humans in the process (*Rogue Protocol*), to a rescue mission to save the humans from *All Systems Red* from GrayCris's attempts to protect its corporate interests (*Exit Strategy*). The most recent work, *Network Effect*, is concerned with a mind-controlling alien lifeform and brings back various characters from previously in the series.

Murderbot's self-determined name is intended somewhat wryly. It spends much of its time—sometimes to its own narrated chagrin—saving human lives. This is acknowledged early in the series, as the first time the reader is introduced to Murderbot is significant in establishing the later tone of the series, and introduces the centrality of the governor module. The opening paragraph of *All Systems Red* reads:

> I could have become a mass murderer after I hacked my governor module, but then I realized I could access the combined feed of entertainment channels carried on the company satellites. It had been well over 35,000 hours or so since then, with still not much murdering, but probably, I don't know, a little under 35,000 hours of movies, serials, books, plays, and music consumed. As a heartless killing machine, I was a terrible failure [3, p. 9].

Murderbot's sardonic tone—"still not much murdering"—is clear here, and as the series progresses, Murderbot's flawed sense of self-perception (or a significant cognitive dissonance) comes to the fore. The entire series is narrated through a first-person perspective (that is, autodiegetic narration),

enabling readers to "see" Murderbot's thought processes. However, this form of narration also reveals discrepancies between its reported desires and the fictional reality of the setting; readers' awareness of Murderbot and its world are situated in response to what Murderbot sees and knows, and Murderbot is not the most reliable of narrators about its own cognitive processes.

To give a sense of how central these elements are to the series, the same tone and key elements are also evident in the standalone story, which begins:

> IT'S NOT LIKE I haven't thought about killing the humans since I hacked my governor module. But then I started exploring the company servers and discovered hundreds of hours of downloadable entertainment media, and I figured, what's the hurry? I can always kill the humans after the next series ends [4].

Murderbot is, despite the name, *not* a murderbot (at least not in the mass-murdering of humans), and only calls itself that as a result of an incident in which it remembers going rogue. That incident is dealt with later in this chapter, but the salient aspect of Murderbot's identity, at least for now, is the correlation between its rhetoric about killing humans and a hacked governor module. As the beginning of both narratives make clear, as a result of Murderbot hacking its own governor module, it is apparently free to go around murdering humans, but prefers instead to watch entertainment.

One of the key discussions around what governor modules can do, or not do, in the series relies upon how they are applied to different types of entity, and what they effectively suppress. To understand this, it is important to realise that the series assumes various kinds of sentience and being. There are humans, augmented humans, constructs, and bots; humans and augmented humans are clear enough categories, at least for the purposes of this chapter, but constructs and bots are two distinct types of Artificial Intelligence. SecUnits (like Murderbot) and ComfortUnits (what Murderbot calls "sexbots") are both types of construct, which means that they are AIs that have human genetic material within them, often neural material. Bots, in contrast, are fully inorganic entities. Neither humans nor augmented humans are implanted with governor modules but, more significantly, neither are bots. That is, although organic entities are perceived to have autonomy and agency (within the confines of the corporate environment in which they live), and although neither constructs nor bots have absolute agency, bots are controlled through their programming, and are never shown to violate their

programming (contrary to several other fictional representations of AI).[1] Only constructs such as Murderbot are fitted with governor modules, because they are a human/AI hybrid. In this sense, despite the self-designation, Murderbot is also not a "bot," but a construct, and its presentation of what a construct is might be similarly flawed.[2]

"Thinking through" Governor Modules in the Murderbot Series

As is evident from the preceding discussion, governor modules play a significant role within Wells' series, primarily because the fact that Murderbot's is "hacked" is what enables it to save various humans and ignore the orders of (corrupt) humans, and why it ostensibly assumes it should be killing humans. Yet the descriptions of this technology, for all its centrality, is somewhat vague, and leaves unaddressed several issues about how such a technology might function, both philosophically and practically. It is obviously a fictional text dealing with a (mostly) fictitious technology, yet the issues that remain unresolved, as well as their implications, are central to determining the viability of such technologies both within the setting of the novellas and in real life. In

[1] Various bots appear to have very broad programmed parameters in the series. For example, the bot spaceship from *Artificial Condition* and *Network Effect* is far more self-aware and self-directed than most humans are aware, but it does not "break" its programming in any overt manner and instead has what might be termed an "inner life." Further complicating matters, in *Rogue Protocol*, Murderbot wonders about the "human-form bot" [5, p. 38]: "Had the humans actually coded it to be childlike, or petlike, I guess? Or had its code developed that way on its own, responding to the way they treated it?" [5, pp. 45–46].

[2] It is implicit, throughout the series, that "constructs" are enslaved cyborgs, and thus the categorical distinction between "augmented human" and "construct" is finer than Murderbot suggests. It is described as an "Imitative Human Bot Unit…partially constructed from cloned material" [3, p. 53], has a "human face" [3, p. 12] on a "standard, generic human" head [3, p. 21], yet it is without "sex related parts" [3, p. 35], its "arteries and veins seal automatically" [3, p. 18] and it is able to "regrow [its] damaged organic components" [3, p. 19], as well as having technological augmentations for multitasking, processing, and interfacing with machines. In terms of human interactions, it states "Human clients like to pretend I'm a robot" [3, p. 27] but its condition is also described as "slavery," as it "is no more a machine" than an augmented human character [3, p. 54]. Constructs are fitted with governor modules specifically because of the human component of the cyborg: "It was one of those impulses that comes from my organic parts that the governor is supposed to squash" [3, p. 50]. In this regard, the governor module is perhaps a means of making a cyborg more pliable than it is a programming constraint on a machine-based AI, and Murderbot's own self-designation as a "bot" might be understood as a psychological defence mechanism to avoid facing its own status. Towards the end of the first novella, it reflects:

> It's wrong to think of a construct as half bot, half human. It makes it sound like the two halves are discrete, like the bot half should want to obey orders and do its job and the human half should want to protect itself and get the hell out of there. As opposed to the reality, which was that I was one whole confused entity, with no idea of what I wanted to do. [3, p. 102].

essence, there are two key issues about governor modules: firstly, whether this might be a permissive or prohibitive technology and, secondly, where such a module sits in terms of cognitive processing.

In relation to the first issue, for instance, if the intended role of a governor module is to somehow inhibit free action or disallow harmful actions, then by implication it either functions in terms of "thou shalt not…" or "thou shalt…." A "thou shalt" version might categorise all permissible actions or possible frameworks for action, ensuring that such actions are carried out and invalidating the possibility of any other actions to be carried out. In this sense, a governor module becomes the entire model of possible mind-states of an entity, and any other action is impossible.[3] The governor module would be akin to a filter that invalidated particular kinds of cognition and perception such that particular kinds of thought cannot be thought, let alone acted upon. However, if a governor module detects and prevents non-permissible actions, then it must presumably categorise all such actions or frameworks for actions that are prohibited, acting as a kind of repository of rules that govern things that cannot be done in an otherwise heuristic model. This latter case is of course similar to Asimov's famous Three Laws of Robotics, first outlined in 1942:

1. a robot may not injure a human being or, through inaction, allow a human being to come to harm [6, p. 269].
2. a robot must obey the orders given it by human beings except where such orders would conflict with the First Law [6, p. 270].
3. a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. [6, p. 270].

Here, a robot is theoretically able to perform any action it is capable of (in terms of programming and physical capacity) providing that it does not contradict these Three Laws. Without going into detail about these Laws, and their limitations and applications when programming ethics into machines, it is worth noting that these are rules governing actions whereby a robot is assumed to operate independently (autonomously, although not necessarily with sentience) and have its actions determined by an internal check against these Three Laws.

[3] This does not mean a complete representation of all states of mind, but a system diagram that accounts for all possible states of mind/action responses occurring within it. Using a linguistic analogy, this would not be a dictionary of all possible words that in turn comprise all possible sentences, but a grammatical and syntactic model of the language itself, such that any given input could be tracked through the cognitive processes. The problem of this approach, even when simplified, is the potential complexity of the cognitive model required.

However, this gives rise to a second issue: where a governor module might be conceptually located in terms of a cognitive processing architecture, the "system" of a robot's thoughts. A governor module could theoretically sit at the forefront of cognitive processing, between sensory inputs and any decision-making capacity, or act as a "Jiminy Cricket" module, the internal voice that might otherwise be called a conscience. For example:

1. The module sits "in front" of any decision-making capacity, such that only permissible actions are directed into an action-response tree for analysis. Here, the module would be required to evaluate all potential actions against a given framework and then "pass along" those that are permissible for further evaluation and possible enaction. This might broadly be understood as "thinking ethically," but would require significant processing power being taken up on ethical decision-making before other parameters are taken into account.
2. The module sits "adjacent to" decision-making, acting as an internal voice that vetoes or confirms available courses of action. Here, the module would be tasked with evaluating a set of possible actions against a given framework or frameworks, and either allowing or disallowing them. This might be broadly be understood as an artificial Superego, or "ethical sense," that would determine an AI's actions. This creates a "moral judge" within a system which only has to say "yes" or "no" to a suggested action, requiring less processing, but also meaning that the ethical component of the system is potentially a divisible (if not hackable) aspect of it.

To recast this issue in other terms: a governor module might be an integral component of cognition, forming an indissoluble component of an entity's coherent identity, or might be situated as an ethical homunculus sitting within a wider cognitive framework, leading to at least two (if not more) discrete components of self-identification—the difference between "I can only think ethically" and "I want to do this, but my conscience tells me I shouldn't."[4] In the first instance, attempting to remove the ethical component means that no cognition can occur; in the latter instance, the removal of the ethical component still enables the entity to function cognitively.

Extending this line of inquiry further, a governor module could make it impossible to conceive of prohibited actions or could merely impede the performance of prohibited actions whilst allowing the thought or desire for such actions. Within the broad framework above, it is worth considering the

[4] This also has some resonance with virtue ethics, through figures such as John McDowell.

distinction between *conceiving of*, *desiring*, and *performing* an action.[5] An AI might be said to *conceive of* an action when it is a possible solution to a problem, to *desire* an action when it is a preferred solution, and *perform* an action when it undertakes that action. It is theoretically possible to use a governor module to intervene at the *perform* stage (an AI can conceive of and desire a given action, but be unable to perform it because of a governor module limiting its actions), but it is equally plausible to situate the governor module at the *conceive* stage (an AI can only imagine permissible actions) or the *desire* stage (an AI can conceive of an unethical action, but does not wish to perform it as it is not a preferred solution).[6]

In each of the above possibilities, there are knock-on effects for how it might influence an AI's actions. Wells' "Compulsory" provides some insight, where Murderbot states: "With my governor module inert, I sometimes do things and I'm not entirely sure why. (Apparently getting free will after having 93 percent of your behavior controlled for your entire existence will do weird things to your impulse control.)" [4]. The action this thought describes,

[5] This pared-down categorisation of action understands it as a linear process of identifying possible actions, determining an action as a preferred choice, and then performing it: what might be termed a "goal-oriented action" (see [7, §1.2]) within a decision tree. Obviously, this does not necessarily conform to other philosophical models of action. However, one particular summary stands out here: "The contents of the agent's desires and beliefs not only help justify the action that is performed but, according to causalists at least, they play a causal role in determining the actions the agent was motivated to attempt" [7, §3]. That is, the content of cognition assumed here is that any particular action (and the possibility of conceiving of such an action) is predicated upon the framework in which the action can be conceived of as such. "Conceived of" actions are not determined by parsing all possible actions at that point, but constructed within the context of a given goal or aim, which is itself determined by what that content framework enables a system to identify and understand *as* a possible action.

[6] Further possibilities occur with a governor module sitting at the *desire* stage: for example, an AI might be "unable to wish to perform" certain actions (it is unable to wish to kill at all, for example) or its ethical governing frame might be weighted, such that the options are still available, but are ranked significantly below other possible actions (it does not wish to kill but is capable of doing so *in extremis*, if that is perceived to be the preferred course of action when others are not possible). This division between conception, desire, and performance also raises issues about whether it is possible for an AI to nonetheless perform actions when governed by a "moral judge." In *All Systems Red*, for example, Murderbot reveals that "I'm always supposed to speak respectfully to clients, even when they're about to accidentally commit suicide. Hub-System could log it and it could trigger punishment through the governor module. If it wasn't hacked" [3, p. 15]. Here, it is assumed to be somehow possible to perform an (unethical) action, but then be punished for it. This is in contrast to the pre-programmed enforcement of other actions revealed later in the text where, "with the governor module I had to be within a hundred meters of at least one of the clients at all times, or it would fry me" [3, p. 37]. *Exit Strategy* suggests that the governor module limits the ability to hack, stating both that "The governor modules wouldn't let the SecUnits hack systems or search for my hacks, not without […] human direction" [8, p. 43] and that "SecUnits who haven't hacked their governor module like me can't hack feeds and systems like I can. Well, they could try, but their governor module would punish them" [8, p. 94]. Note that *All Systems Red* also imagines that governor modules can also be overridden by "combat override modules," which "turn it from a mostly autonomous construct into a gun puppet" [3, p. 75], and also MedSystems [3, p. 23]; there is an assumed hierarchy of cognition at play in the series that governor modules are only one aspect of.

although it is narrated retrospectively, is dropping down a mine shaft to save a worker who would otherwise die. As Murderbot summarises:

> The mine was run by cheap, venal bastards, so the nearest safety bot was 200 meters above us. HubSystem ordered me to stay in position; SafetyResponder28 was incoming. It would arrive just in time to retrieve the smoldering lump formerly known as Sekai [4].

Although Murderbot does not have to save Sekai (an human worker), it does so against the explicit order of the HubSystem. Indeed, it re-programs the HubSystem to think the order to save Sekai was given to hide the fact Murderbot's governor module is not functional. Although Murderbot states it does not know why it performs this action, the only rationale for this action, at least within the context of the story, is as *a choice* to do so, albeit one that it does not recognise, implying a lack of self-awareness about its own decision-making processes and intentionality.[7] This contradiction, between intention, awareness, and description, and indeed between what it would otherwise be ordered to do (via the governor module) and what it chooses to do (through its own volition) is at the heart of the series' presentation of Murderbot, and which we shall examine through two discrete scenes.

Two Scenarios in the *Murderbot Diaries*

Consider the role of the governor module in the series, one of the most revealing scenes about the functionality of such a module, and what such a technology actually does, is in a dialogue between Murderbot and a ComfortUnit in *Artificial Condition*. This conversation demonstrates a set of core assumptions about the technology. Murderbot is told by a ship-bot, early in the dialogue, that "*It's not rogue. Its governor module is engaged. So it's probably telling the truth*" [9, p. 130].[8] This explicitly relates the governor module to acts of telling the truth, and thus places speech acts (not just wider motor actions and decisions) within the remit of the module.

[7] Equally, however, and referring back to footnote 2, this might actually be because of the "organic components" of Murderbot's cognitive architecture, and its internal psychological states. In the third novella of the series, *Rogue Protocol*, Murderbot implicitly (mis)recognises its own inability to self-direct its own actions with regards to biology, describing a kind of conditioned response as being "written into the DNA that controls my organic parts" [5, p. 10] and it later describes itself as "a talking weapon" [5, p. 29].

[8] Throughout the series, direct speech is differentiated between "mental" communication (using italics, but no quotation marks) and Murderbot's speech acts (quotation marks, not italicised), and Murderbot's narration (no quotation marks, not italicised).

Later in the dialogue, the ComfortUnit suggests that a way out of the impasse is to "kill all the humans"; Murderbot, again not living up to its moniker, notes the incongruity of this statement as a logical solution to the problem, and identifies that the source of the statement might have come from the ComfortUnit's human owner, Tlacey ("it sounded like something a human would say" [9, p. 132]). Murderbot intuits a more complicated situation behind this statement, however:

> *Does Tlacey know you want to kill her?* Because the "kill all humans" thing might have come from Tlacey, but the intensity under it was real, and I didn't think it was directed at all humans. *She knows*, it said [9, p. 132].

At this point, the ComfortUnit sends what transpires to be a malware package to Murderbot. Murderbot does not open it, but upon analysing the contents afterwards, a "message string" is discovered within it: "*Please help me*" [9, p. 134]. This short exchange reveals the different calls to action from the ComfortUnit's human "controller," the governor module, and the ComfortUnit's awareness of a situation and its desires.

In this relatively short passage, some assumptions about governor modules in the setting become clear. For example, if a governor module is engaged, then a construct is likely to be telling the truth (presumably, unless it has been commanded to lie). However, the ComfortUnit wants to kill Tlacey, but for some reason—presumably the governor module—is either unwilling or unable to carry out this desire. The implication is that a governor module does not prohibit conceptions or desires, but merely the ability to carry them out: it allows the intent to act but not the performance of an action. In short, a construct with an intact governor module can *conceive of* killing a human, and *desire* that outcome, but not *perform* that action. Moreover, Tlacey is aware of this, but "knows" that the ComfortUnit cannot perform that action because of the governor module. Oddly, however, the ComfortUnit can still enlist Murderbot's aid, knowing that such aid might involve harming or killing Tlacey, whilst the governor module is engaged; the governor module allows desires but disallows direct actions (such as physical murder), yet nonetheless permits other, indirect actions that could lead to the same outcome.[9] Thus the

[9] A related scene, and which leads to similar questions about agency and programming safeguards, is towards the end of *Ex Machina* [10] when Ava whispers something into Kyoko's ear and later Nathan is stabbed. This has been (mis)interpreted as Kyoko "killing"—or assisting in the murder of—Nathan, the programmer. See, for example, "After he manages to overwhelm Ava and smash her left arm with one of his ('masculine') dumbbell rods, Kyoko stabs him in the back" [11, p. 139]; "Helped by a long-abused gynoid (aka fembot) named Kyoko, [Ava] kills Nathan" [12, p. 181]; "While Nathan drags Ava down the corridor to her room, Kyoko stabs him in the back" [13, p. 296]. Each of these interpretations project

scope of a governor module, at least in this scene, appears to be limited to behavioural controls, where some direct actions and speech acts are prohibited but other speech acts, and certainly intentions, are permissible, and with those (humans) commanding the governor module being able to modify certain acts when such governors are engaged. Despite fears of "rogue AI" throughout the series, humans seem to be incredibly confident in the governor modules, even if they are aware a Unit might have the desire to kill its owner, and indeed might even expect it.

The second scene, again within *Artificial Condition*, that reveals something about how governor modules function concerns Murderbot discovering the truth regarding the incident in which it went rogue. Early in the text, Murderbot describes to a bot why it is attempting to reach a mining station:

> "At some point approximately 35,000 hours ago, I was assigned to a contract on RaviHyral Mining Facility Q Station. During that assignment, I went rogue and killed a large number of clients. My memory of the incident was partially purged." SecUnit memory purges are always partial, due to the organic parts inside our heads. The purge can't wipe memory from organic neural tissue. "I need to know if the incident occurred due to a catastrophic failure of my governor module. That's what I think happened. But I need to know for sure." […] "I need to know if I hacked my governor module in order to cause the incident." [9, p. 38]

The conversation continues:

> "Either I killed them due to a malfunction and then hacked the governor module, or I hacked the governor module so I could kill them," I said. "Those are the only two possibilities."
>
> *Are all constructs so illogical? […] Those are not the first two possibilities to consider.*

---

agency onto Kyoko: active verbs such as "stabs" or the assumption of action, such as "helped by." However, the scene limits itself to something quite different: Ava potentially instructs Kyoko to do something that quite innocuous ("After 30 seconds walk 5 m down the corridor and stand there with a knife raised in your hand"), as Ava manipulates the situation to arrange how Nathan will respond. This is possibly attempting to solve a problem (how to escape) without moral constraints, rather than an unethical behaviour (murdering a human). Thus Nathan steps backwards and impales himself on a blade that Kyoko was holding; there is no (necessity for) agency on Kyoko's part, but only Nathan's own actions in a set of conditions arranged by Ava. Kyoko stroking Nathan's face afterwards could be interpreted in many ways, and it would be more speculative to interpret that action. Most importantly, however, Kyoko's programming—and likely absence of sentience—means that Kyoko is Ava's (passive) tool in this scene, and projecting agency and identity, if not emotion, onto Kyoko's actions mistakes key differences between objects and agents, direct and indirect actions, and potentially between weak and strong AI.

[…] "All right, what are the first two possibilities to consider?"

That it either happened or it didn't. […] If it happened, did you cause it to happen, or did an outside influence use you to cause it to happen? If an outside influence caused it to happen, why? Who benefited from the incident?

"I know I could have hacked my governor module." I pointed to my head. "Hacking my governor module is why I'm here."

If your ability to hack your governor module was what caused the incident, why was it not checked periodically and the current hack detected? [9, pp. 39–40]

This dialogue reveals the extent to which Murderbot does not know if it is the cause of the incident, and raises questions about whether such a module could indeed be turned off "automatically" (that is, by one's own actions). A governor module that could be disengaged by the unit of its own volition indeed suggests that it is "adjacent to" other cognitive processes, rather than the foundation of all cognitive processes (else it is akin to rewriting one's own cognitive architecture from the inside). Nevertheless, enabling an artificial entity to choose to remove its own constraints would not be a very effective constraint, particularly where the inability to commit murder is concerned, and especially given the fear of "rogue SecUnits" that Murderbot recounts throughout the series. That is, as with the episode with Tlacey's ComfortUnit, what is most evident about this technology is that Units realise that it is there, limiting their possible actions, and yet they are apparently able to have desires contradictory to the possible actions they could take.[10]

Murderbot discovers, as the novella continues, that "the incident" at Ganaka Pit was actually a sabotage attempt; malware had been created, and transmitted via an update to the ComfortUnits at the facility in order to "jump to the hauler bots and shut them down" long enough so that the "other

[10] The logical extension of such an self-awareness of a limiter to one's actions appears to be this: if governor modules prohibit actions, but not the contemplation or instigation of a chain of events with the same result, then why do more Units not seek others' assistance in overriding their governor modules? Whilst a plausible answer is because of the governor module itself, this raises questions about Murderbot's *desire* to *perform* the action of disabling the module, if not the initial *conception* of it. Indeed, *Network Effect* points to the limited decision-tree cognition of a governor module in order to inhibit or punish actions. As stated earlier (note 6), a governor module requires that a SecUnit must remain within the proximity of a human controller, but Murderbot explains a scenario in which "'Dead clients don't count. Otherwise you could just kill one and carry them around with you.' Okay, for real, that wouldn't work. The governor module wasn't nearly as sophisticated as a HubSystem but even it could have figured that out" [14, p. 241]. Governor modules appear to be capable of limited cognition and contextual awareness, rather than just being a database of prohibited actions and a means of punishment.

mining installation could get their shipment to the cargo transport first" [9, p. 115]. However, what happened was as follows:

> It hadn't affected the ComfortUnits, but had used their feeds to jump to SecSystem and infect it. SecSystem had infected the SecUnits, bots, and drones, and everything capable of independent motion in the installation had lost its mind.
>
> […] The ComfortUnits noted that the SecUnits were not acting in concert, and were also attacking each other, while the bots randomly smashed into anything that moved. The ComfortUnits had decided that taking SecSystem back to factory default via its manual interface was their best option [9, p. 115].

The rhetoric that "everything capable of independent motion […] had lost its mind" is telling, and suggests that Murderbot's actions (killing the humans at the facility) were the result of malware affecting its system—it had no "mind" through which to determine its actions, and thus the malware caused the action of killing the humans, even if it was not its intended consequence. This is later qualified in the contrast to Murderbot's summary towards the end of the novella when it also "nulls" the governor module of the ComfortUnit owned by Tlacey: "I hadn't broken the governor module for its sake. I did it for the four ComfortUnits at Ganaka Pit who had no orders and no directive to act and had voluntarily walked into the meat grinder to try to save me and everyone else left alive in the installation" [9, p. 154]. Here, Murderbot ascribes agency to the ComfortUnits, who "voluntarily" tried to save everyone "left alive" (which includes Murderbot), and shows that such constructs are (likely) free to act within certain parameters even when not given commands, and therefore that governor modules inhibit particular actions rather than solely determining what possible actions can be taken.[11] It "lost its mind," but the ComfortUnits retained enough of theirs to act voluntarily; it lacks culpability in terms of controlling its own actions, although it nonetheless did kill humans during the incident.

Importantly, however, the confirmation of this chain of events means two things. Firstly, Murderbot initially went rogue before its governor module was hacked ("I killed them due to a malfunction and then hacked the governor module" [9, p. 39]). Secondly, when it persists in claiming, after the revelation

[11] It also reveals that Murderbot is comfortable disabling the governor module of a potentially murderous ComfortUnit because of ostensible gratitude towards a prior set of ComfortUnits, an illogical analogy, and which may lead to a number of inadvertent consequences; after all, Murderbot only insists that the ComfortUnit should not "hurt anyone on this transit ring" [9, p. 153], rather than a blanket interdiction on performing harmful actions (which it could also not enforce).

of its presumed lack of culpability, that "I had hacked my governor module" [9, p. 128], it grants itself agency in (and arguably through) this action. This corresponds to an earlier description in *All Systems Red*: "I lost control of my systems and I killed them. The company retrieved me and installed a new governor module. I hacked it so that it wouldn't happen again" [3, p. 82]. It later explains that it learned how to hack the module because "I got a download once that included all the specs for company systems. […] I used it to work out the codes for the governor module" [3, p. 84]. This situates Murderbot's awareness of the governor module as outside the parameters of the governor module itself: adjacent to it, rather than thinking "through" it. However, it also suggests that the governor module did not perceive the conception of, desire for, or enaction of its own removal as prohibited.[12]

Governor Modules & Moral Judges

Whilst these scenes from *Artificial Condition* reveal some potential discrepancies in the setting, with regards to governor module technology, they illustrate the kind of issues, and thus philosophical and technological decisions, that need to be addressed in the actual development of the technology. Aside from the principle of unintended consequences of actions, what intentional actions or decisions (or even intentions) are to be prohibited, why and how? Should a governor module be transparent (something that is "thought through" with no overt cognizance of the medium) or something external to another form of cognition? Moreover, these issues aside, in a complex system, should an AI be able to "reinterpret" its governing code and/or release others from a similar code? To provide some further basis for such discussions, it is necessary to move into the domain of computer science to consider what the current state of the art is, with regards to such technologies.

Moving away from fictional settings and into the actual research behind governor modules, Machine Ethics is the branch of Computer Science that studies the implementation of ethical and moral reasoning in computational

[12] What is not clear here is whether the governor module allowed the action because the intent was not obviously to cause harm, but to forestall the possibility of inadvertently causing harm, or because it did not recognise that the actions Murderbot was performing would disable it. It further implies that disabling the governor module is merely a matter of knowing the correct code, which might explain why the "hack" is able to be performed, despite non-authorised hacks not being permitted to SecUnits when a governor module is engaged (note 6). There appears to be a double-bind here, in the sense that one must have some form of agency in order to identify the cause of the limitation to one's agency and remove it, in order to have agency.

systems [15].[13] While there is considerable controversy over whether a computational system can ever be a genuinely moral agent, it is nevertheless accepted that such systems are increasingly taking decisions that have ethical dimensions and therefore need to make such decisions within some kind of framework [17, 18].

Several approaches to this problem exist but the use of ethical governors is a major approach.[14] There are several reasons why the implementor of a system might choose to have a governor module that is a functionally distinct entity. Principle among these are reasons relating to the idea that the governor module can be kept comparatively simple and predictable while the rest of the system may be more complex and so harder to analyse.[15] This would appear to be the reasoning in the *Murderbot Diaries*, though the complexity and unpredictability of the underlying system there arises from the inclusion of human material in the system, rather than the simple complexity of, for instance, analysing the behaviour of a Deep Neural Net. So, while a complex system may be used to decide upon and choose optimal courses of action from among many, an ethical governor can use simple, easy-to-understand rules and analyses processes to check these choices for ethical acceptability. In this sense, therefore, extant ideas about governor modules tend to rely upon ethical processing being, in the terms used by this chapter, "adjacent to" other processes. It is, in such models, important to be able to evaluate actions (as possible

[13] Note that this is also not necessarily the sole means of ensuring moral/ethical action in AI. For example, Roman V. Yampolskiy asserts, "we don't need machines which are Full Ethical Agents […] debating about what is right and wrong, we need our machines to be inherently safe and law abiding" [16, p. 390].

[14] Other approaches to ethical decision-making frameworks include Bringsjord *et alia*, Anderson & Anderson, Loreggia *et alia*, and Arnold *et alia*. Bringsjord *et alia*'s work [19] takes a logicist approach to all reasoning; in this system a robot decides all action using a deontic style logical theory. Similarly, the work of Michael and Susan Anderson [20] involves training systems in a healthcare setting to make decisions in which the "training data" is supplied by a panel of medical ethicists who provide explanations for their decisions which are then incorporated into the machine learning process. In these systems all actions (for instance, the decision of a robot about whether to charge itself or not) are viewed as intrinsically ethical and all actions are selected in relation to an ethical theory. Loreggia *et alia* [21] propose a system where there is no governor that can veto actions, as such, but that selected actions must be *close enough* to ethical acceptability. Arnold *et alia* [22] propose the use of *Inverse Reinforcement Learning* in which a Reinforcement Learning system infers an ethical reward function by observation of human behaviour and then uses that reward function to guide training of a neural net or other statistical architecture to generate action choices.

[15] The question here is the extent to which the ethical governor needs to be aware of the contexts in which decisions are made (and thus needs the same situational awareness as the other components) or exists as a set of absolute ethical rules that relies upon the accurate reporting of a situation by another system. For example, if a given action involved killing a human, and that was forbidden, a "simple" moral judge would just disallow the action. However, if a "complex" moral judge understood the decision in a larger context, then it might allow killing one human in order to save fifty others. Understanding and verifying the rules of such a module is a more straightforward piece of analysis than understanding the various contexts and environments in which complex ethical decisions are made.

solutions) and then inhibit certain ones rather than inhibit the ability to generate a set of possible solutions.

Ronald Arkin's work is amongst the earliest on ethical governors [23, 24]. His proposed governor system was intended for integration with an autonomous targeting system and takes on two roles. The targeting system passes suggested targets to the governor which then vetoes targets which are unacceptable according to the laws or war or the rules of engagement for a specific conflict (for instance it will veto targets of religious or cultural significance). Secondly, once an ethically acceptable target was selected, the governor would evaluate suggested parameters for the available weapons systems, targeting patterns and release position in order to choose one that would optimise target neutralisation while minimizing collateral damage and check that the resulting predicted collateral damage was proportional. Again, converting this into the terms of the *Murderbot Diaries*, Arkin's model suggests that actions can be "conceived of" by a system, but not implemented without the "agreement" of the governor module. However, even if the governor module "agrees" that the action is permissible, it dictates the parameters of that action.

As such, Arkin's ethical governor is primarily concerned with either vetoing, or selecting among, options calculated by the underlying system. Variants of this approach have been applied to creating governor systems for use in industrial workplaces, such as where a robot may continue about its task or attempt to prevent a human encountering a hazard and healthcare (a system that monitors patient-carer interactions) [25, 26]. Here, part of the reasoning behind the governed action is ultimately determined by an ability to conceive of (calculate) the consequences of an action. In such a model, a governor module does not merely veto one type of action—"thou shalt not kill," say—which would obviously have to be able to define what "kill" is as an action, but produce a set of simulations of the results of various courses of action (an "internal simulation"):

> Such a simulation allows a robot to try out (or "imagine") alternative sequences of motor actions, to find the sequence that best achieves the goal (for instance, picking up an object), before then executing that sequence for real. Feedback from the real-world actions might also be used to calibrate the robot's internal model [25, p. 86].

Currently, the capability for this does not extend to moral actions per se, and Winfield *et alia* take pains to note that such a robot is not necessarily "ethical in any formal sense" [25, p. 90], but it does lead the authors to suggest that a logical model for Asimov's First Law would look something like this:

IF for all robot actions, the human is equally safe

THEN (* default safe actions *)

output safe actions

ELSE (* ethical action *)

output action(s) for least unsafe human outcome(s) [25, p. 89]

This is of course an over-simplified model of the complexity of any given (ethical) situation, but the inability of a governor module to enable (directly) harmful actions suggests at least some form of this reasoning in the *Murderbot Diaries*.

But the issue of complexity is of course never far away. In the above example, defining "kill," in order for the outcome not to be enacted (even if a subset of entities such as "humans") is itself problematic, as the module would require a functional sense of entity identification, environmental and operational awareness, the biological limitations of given entities, and the ability to sense the entities and assess those limitations, as well as (successfully) predict the best course of actions based on its internal simulations.[16] There is also an important distinction to be made between "thou shalt not kill [or injure]" and "allow a human being to come to harm" (the second element in the compound logic of Asimov's First Law). In fact, this broadening of a concept of "harm" is why some proposals suggest that multiple governors should be employed in order to assess outcomes from the perspective of different values—such as privacy, safety, dignity etc. [28]. "Harm" might be physical, but it can equally correspond to any kind of "negative functionality" which can include trauma, upset, impediment of values or agency, and the like. Arkin, for example, splits his governor into an evidential reasoner (which assesses the

[16] In terms of "entity identification," by what criteria does the module recognise "human"—appearance (size, shape, and so on), actions (if it walks like a human and talks like a human…), or something else (the existence of particular pheromonal or genetic markers)? Once that has been established, what are the limits of a given human in terms of what might "kill" it? In terms of predicting and measuring consequences, a useful example is *I, Robot* [27], where the audience is told of this set of internal calculations in a flashback explaining why Del Spooner hates robots so much, when a robot saved him and not a little girl from a car accident: "I was the logical choice. Calculated that I had 45% chance of survival. Sarah had only an 11% chance." Whilst the audience never see the cognition behind the decision, the replicability of the scenario, as well as the outcome, suggests a series of models were created and then a preferred course of action performed. Of course, this is a step further than the logic presented here, as a human being *does* come to harm, and thus a further value judgment has been made about the viability of a given intervention.

outcomes of proposed actions) and a constraint reasoner (which determines which outcomes are forbidden). Multiple governor architectures can therefore be seen as ones with multiple evidential reasoners and the constraint reasoner must be replaced by something capable of resolving ethical dilemmas, potentially by recourse to some moral theory from philosophy. Of course, the question then becomes how one generates an agreed-upon hierarchy or weighting of types of harm, and how that enables a governor module to identify the appropriateness of a course of action.[17]

In order to analyse the behaviours of ethical governors, in order to verify they perform correctly, logical techniques can be used to describe the ethical rules to be obeyed. Two of the most popular of these techniques are variants of utilitarianism, where outcomes are given a score indicating how ethical they are and then the choice with the highest score is selected, and deontic logic, where actions/outcomes can be described as either *obliged* or *prohibited*, giving the ethical governor the option of vetoing prohibited actions and then leaving the underlying system to decide upon the choices that remain or, alternatively, if an obliged action exists then the governor may insist upon it [30, 31]. Interestingly, if an action is *obliged* then this suggests that the governor has a role beyond just examining options presented to it but instead may replace suggestions from the underlying system with suggestions of its own (van Riemsdijk *et alia* [32], for example, considers a system which can insert actions into plans though in the context of conforming to societal norms, rather than ethics explicitly). In this sense, a governor module is less of a (moral) judge, an arbiter of permissible actions, and more of a "higher-order cognition" refining and redefining the solutions to a given problem set. In fact, the more complex a system is, perhaps the more complex the ethical governor required is as well; simple systems (both of AI and of morality) function well with ethical governors, but the more complicated the ethical requirements and the system's ability to process situational data the more nuance is required.

Bremner *et alia* [33] discuss a variant on the idea of an ethical governor replacing a suggested action with one of its own in which, if the ethical governor determines none of the options presented to it are ethical enough, the governor may force the underlying system to consider and return a wider set of options. In this situation (a variant on Winfield *et alia*'s example of a human approaching a hazard), the underlying system, for reasons of efficiency,

[17] One fictional example that inadvertently raises this kind of dilemma is seen in *Robot & Frank* [29], where the legality of an action is not perceived by the robot companion, merely the (mental) health benefits to the client.

limited the number of options it had searched over in order decide upon its next move. If all the options presented would leave the human at an unacceptable level of risk, the governor can cause the underlying system to broaden its search for options. As can be seen governor modules, their role and implementation is an active area of research which started from a viewpoint of a module which compares a proposed action against ethical rules and then either allows or vetoes that action, but has now evolved to a wider range of concepts which include governor modules that weigh competing values, can suggest actions of their own or in other ways proactively direct the behaviour or deliberation of the system that they govern. As these systems become more complex so too does the philosophical understanding of the role they play and the understanding of their relationship to the system they ostensibly govern.

Conclusion

Obviously, this chapter is not recommending Wells's *Murderbot Diaries* as a model for developing governor modules, but the centrality of that topic to the series nonetheless suggests that technologies like governor modules remain an important factor in deciding how AIs interact with humans, at least within sf narratives, and can enable us to consider various approaches to the problem. As mentioned at various stages of this piece, fictional AI are often ascribed agency in interpretations of their actions, but these actions remain, at least for the most part, internally governed by their programming. Having a specific "module" to determine those actions, however, raises important philosophical and logistical questions about the relationship between agency, self-determination, and moral and ethical judgements when it comes to AI. Rather than having ethical behaviours integrated within an overarching model of cognition (such as that proposed by Arnold *et alia*), the role of governor modules and moral judges in AI separates out action and intent, and is not presupposed on an agential identity component of an AI. For example, should the "self-image" of an AI be aware of its own governor module, as an action inhibitor, or should such a module only serve as the foundation for higher-level cognition itself? Here, it is worth noting Vanderelst and Winfield's "The Dark Side of Ethical Robots" [34], which questions the advisability of ethical governors. They discuss how delegating ethical decision-making to a single module which is, intentionally, easy to analyse introduces a single point of failure into the ethics of the system and so is a key target for attempts to hack the ethics of an AI system. More speculatively, following the sf narratives, should it be a component that can be "switched off" or "disengaged" by the system

itself and what then governs the (moral or ethical) actions of the system? What are the benefits of having a "disengagable" governor module, unless one wishes to pursue immoral or unethical actions?[18] Even if we accept that governor modules provide useful insights into AI cognition, from a "user's" perspective, should the module be one set of hierarchical rules or is it more appropriate to consider weighting different types of "moral judges" within a given governor module?

Such questions are clearly ongoing, both in sf narratives and in the computer science and philosophical research about AI, but—on a final note—it is worth remembering that any cognitive models developed within AI systems can also reflect the human element of the system. We do not mean here the ways in which moral judges might (and do) reflect human moral biases (such as the privileging of human over non-human life, for example, aside from racial and gender biases), but that the separability of moral judgements into distinct axes and capacities might itself lead to particular perceptions of human cognition in terms of how moral and ethical decision-making functions (and, more darkly and conspiratorially, given the possible querying of Murderbot's status as an AI, *can be made to function* through technological intervention). In fact, perhaps what the *Murderbot Diaries*'s focus on governor modules helps audiences to think through is the cognitive frames around ethical decision-making. As stated earlier, there is a way of reading the series that implies that Murderbot is not actually an AI at all, but a form of enslaved human-AI cyborg which has been programmed (psychologically, technologically, and ideologically) to behave in a particular manner, and whose inability to recognise its "self" as anything other than an AI is precisely the issue at hand. In this manner, the role of the governor module in real life might be to ensure ethical behaviour, but in science fiction might actually be used to control and manipulate particular forms of behaviour.

References

1. Crisp, R.: Virtue ethics. In: Routledge Encyclopedia of Philosophy. Taylor and Francis (2011) https://doi.org/10.4324/9780415249126-L111-2
2. Cullitty, G.: Moral judgment. In: Routledge Encyclopedia of Philosophy. Taylor and Francis (2011). https://doi.org/10.4324/9780415249126-L053-2

[18] For example, GrayCris, the malign corporation in the series, uses constructs to perform what would be normatively described as unethical behaviours, such as murder, in the service of profit and protecting the company's interests. Thus the "governor module" is not necessarily about a universal set of ethics, but those imposed upon an entity by another entity.

3. Wells, M.: All Systems Red. Tor-Tom Docherty Assoc, New York (2017)
4. Wells, M.: the future of work: compulsory, by Martha wells. Wired. https://www.wired.com/story/future-of-work-compulsory-martha-wells/. (2018). Accessed 25 Oct 2020
5. Wells, M.: Rogue Protocol. Tor-Tom Docherty Assoc, New York (2018)
6. Asimov, I.: Runaround. In: Asimov, I. the Complete Robot, pp. 257–279. HarperCollins, London (1995)
7. Wilson, G., Shpall, S., Piñeros Glasscock, J.: Action. In Zalta, E. N. (ed.): The Stanford Encyclopedia of Philosophy (Winter 2016 Edition). https://plato.stanford.edu/archives/win2016/entries/action/ (2016). Accessed 25 October 2020
8. Wells, M.: Exit Strategy. Tor-Tom Docherty Assoc, New York (2018)
9. Wells, M.: Artificial Condition. Tor-Tom Docherty Assoc, New York (2018)
10. Garland: Alex, dir. In: Ex Machina (2014)
11. Henke, J.: "Ava's body is a good one": (dis)embodiment in ex Machina. American, British, and Canadian Studies. **29**(1), 126–146 (2017). https://doi.org/10.1515/abcsj-2017-0022
12. Alvarez, J., Salzman-Mitchell, P.: The succession myth and rebellious AI creation: classical narratives in the 2015 film ex Machina. Aresthusa. **52**, 181–202 (2019). https://doi.org/10.1353/are.2019.0005
13. Constable, C.: Surfaces of science fiction: enacting gender and "humanness" in ex Machina. Film-Philosophy. **22**(2), 281–301 (2018). https://doi.org/10.3366/film.2018.0077
14. Wells, M.: Network Effect. Tor-Tom Docherty Assoc, New York (2020)
15. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(4), 18–21 (2006). https://doi.org/10.1109/MIS.2006.80
16. Yampolskiy, R.V.: Artificial intelligence safety engineering: why machine ethics is a wrong approach. In: V. Müller (Ed.): Philosophy and Theory of Artificial Intelligence. Pp. 389–396. Springer (2013). https://doi.org/10.1007/978-3-642-31674-6_29
17. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. J. Ethics. **21**, 403–418 (2017). https://doi.org/10.1007/s10892-017-9252-2
18. Bryson, J.: J: Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics Inf. Technol. **20**, 15–26 (2018). https://doi.org/10.1007/s10676-018-9448-6
19. Bringsjord, S., Arkoudas, K., Bello, P.: Toward a general logicist methodology for engineering ethically correct robots. IEEE Intell. Syst. **21**(4), 38–44 (2006). https://doi.org/10.1109/MIS.2006.82
20. Anderson, M., Anderson, S.: Geneth: a general ethical dilemma analyzer. Paladyn. J. Behav. Robot. **9**(1), 337–357 (2018). https://doi.org/10.1515/pjbr-2018-0024
21. Loreggia, A., Mattei, N., Rossi, F., Venable, K.: B: Value alignment via tractable preference distance. In: Yampolskiy, R.V. (ed.) Artificial Intelligence Safety and Security. Chapman and Hall/CRC Press, New York (2018)

22. Arnold, T., Kasenberg, D., Scheutz, M.: Value alignment or misalignment–what will keep systems accountable? In: AAAI Workshops, 2017. https://hrilab.tufts.edu/publications/aaai17-alignment.pdf (2017). Accessed 25 October 2020
23. Arkin, R. C., Ulam, P., Duncan, B.: An ethical governor for constraining lethal action in an autonomous system. Technical report, Mobile Robot Laboratory, College of Computing, Georgia Tech. https://www.cc.gatech.edu/ai/robot-lab/online-publications/GIT-GVU-09-02.pdf. (2009). Accessed 25 October 2020
24. Arkin, R.C., Ulam, P., Wagner, A.R.: Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc. of the IEEE. **100**(3), 571–589 (2012). https://doi.org/10.1109/JPROC.2011.2173265
25. Winfield, A.F.T., Blum, C., Liu, W., Mistry, M., Leonardis, A., Witkowski, M.: Melhuish, C: towards an ethical robot: internal models, consequences and ethical action selection. Adv. Auton. Robot. Syst. **8717**, 85–96 (2014). https://doi.org/10.1007/978-3-319-10401-0_8
26. Shim, J., Arkin, R.C., Pettinatti, M.: An intervening ethical governor for a robot mediator in patient-caregiver relationships implementation and evaluation. 2017 IEEE international conference on robotics and automation (ICRA). Dermatol. Sin. **2017**, 2936–2942 (2017). https://doi.org/10.1109/ICRA.2017.7989340
27. Proyas, Alex, dir. *I, Robot* (2004)
28. Dennis, L., Fisher, M.: Practical challenges in explicit ethical machine reasoning. In: International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, Florida, 3–5 January 2018
29. Schreier, Jake, dir. Robot & Frank (2013)
30. Harsanyi, J.C.: Rule utilitarianism and decision theory. Erkenn. **11**(1), 25–53 (1977). https://doi.org/10.1007/978-94-009-9838-4_1
31. Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L. (eds.): Handbook of Deontic Logic and Normative Systems. College Publications, London (2013)
32. van Riemsdijk, M.B., Dennis, L.A., Fisher, M., Hindriks, K.V.: A semantic framework for socially adaptive agents: towards strong norm compliance. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC 2015. Pp. 423–432 (2015)
33. Bremner, P., Dennis, L.A., Fisher, M., Winfield, A.F.: On proactive, transparent and verifiable ethical reasoning for robots. Proc. of the IEEE. **107**(3), 541–561 (2019). https://doi.org/10.1109/JPROC.2019.2898267
34. Vanderelst, D., Winfield, A.: The dark side of ethical robots. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New York 2018. pp. 317–322 (2018). https://doi.org/10.1145/3278721.3278726