# Robots, institutional roles and joint action
# some key ethical issues

Miller, Seumas

**DOI**
10.1007/s10676-024-09816-z

**Publication date**
2025

**Document Version**
Final published version

**Published in**
Ethics and Information Technology

Ethics and Information Technology (2025) 27:10
https://doi.org/10.1007/s10676-024-09816-z

ORIGINAL PAPER

# Robots, institutional roles and joint action: some key ethical issues

**Seumas Miller**[1,2,3]

Accepted: 4 December 2024


**Abstract**
In this article, firstly, cooperative interaction between robots and humans is discussed; specifically, the possibility of human/robot joint action and (relatedly) the possibility of robots occupying institutional roles alongside humans. The discussion makes use of concepts developed in social ontology. Secondly, certain key moral (or ethical—these terms are used interchangeably here) issues arising from this cooperative action are discussed, specifically issues that arise from robots performing (including qua role occupants) morally significant actions jointly with humans. Such morally significant human/robot joint actions, supposing they exist, could potentially range from humans and robots jointly caring for the infirm through to jointly killing enemy combatants.

**Keywords** Robots · Autonomous systems · Joint action · Organisational action · Social institutions · Moral properties · Moral responsibility

## Introduction

In this article, firstly, cooperative interaction between robots and humans is discussed; specifically, the possibility of human/robot joint action and (relatedly) the possibility of robots occupying institutional roles alongside humans. Such roles might potentially range from ones involving the performance of simple manual tasks, e.g., assisting human to lift furniture onto trucks or hospital patients onto their beds, through to ones involving complex tasks, e.g., 'fighting' a war under the direction of human combatants as in the case of a predator drone that identifies, tracks and fires weaponry at enemy combatants. The discussion makes use of concepts developed in social ontology, notably those of joint action and social institution. However, as is to be expected, there are multiple, conflicting theories of these concepts in the academic literature and, therefore, space does not permit the elaboration and adjudication of all or even most of these theories in what is essentially an article in applied, as opposed to purely theoretical, philosophy. Accordingly, I have had to be selective with respect to these theories. In doing so I have opted for the so-called collective end theory of joint action (Miller, 1992, 2001, 2007) and the normative-teleological theory of social institutions (Miller, 2010) both of which I have elaborated and defended in detail elsewhere.

Secondly, certain key moral (or ethical—these terms are used interchangeably here) issues arising from this cooperative action are discussed, specifically issues that arise from the possibility of robots performing (including qua role occupants) morally significant actions jointly with humans. Such morally significant human/robot joint actions, supposing they exist, could potentially range from humans and robots jointly caring for the infirm through to jointly killing enemy combatants. Moreover, an issue arises with respect to human/robot 'joint' actions in which one of the participants has some degree of institutional authority over the other participant(s), e.g., as a commanding officer has over their subordinates or a manager has over their lower echelon employees. Let us suppose that such actions can be, in some instances, a species of joint action, notwithstanding the existence and influence of some form of institutional authority. Here if A has institutional authority over B, then A has morally and institutionally legitimate power over B (with respect to some range of tasks). A question now arises as to whether the individual contributory 'actions' of the robot(s) in such joint actions must always be under the control (authority?) of one or other of the humans. In

✉ Seumas Miller
semiller@csu.edu.au

[1] Charles Sturt University, Canberra, Australia

[2] University of , Oxford, UK

[3] TU Delft, Delft, Netherlands

Published online: 21 December 2024

short, is the concept of institutional authority, as opposed to causal control, potentially applicable to robots? If not, then robots cannot occupy institutional roles in any deep sense since they are immune to institutional authority. Further, the concept of institutional authority entails institutional rights and duties, but such rights and duties are frequently, if not invariably, dependent on *moral* right and duties (Miller, 2010). Accordingly, if robots are not moral agents, then this is in and of itself a barrier to their possession of institutional rights and duties—and, therefore, occupancy of institutional roles.

Overall, the discussion in this article of human/robot joint action in general, and of such action in institutional settings in particular, is deflationary. This is in large part because robots are not moral agents (Sharkey, 2012; Sparrow, 2021). Here I note that new and emerging technologies are developing rapidly making it difficult to predict what might or might not be able to be achieved in the future. For this reason, I have adopted the conservative stance of restricting my ethical analysis to existing robots and robot technology rather than those that are the object of speculation or even informed extrapolation. This has the consequence that what I might regard as a limitation of robots might be thought by others to be merely a limitation of currently existing robots and robot technology. However, since, by definition, future robots and robot technology are not yet with us the general claim that current limitations will be overcome cannot at this time be confirmed or disconfirmed on the basis of empirical evidence. This is especially the case when it comes to highly ambitious, largely speculative claims, such as the claim that robots could become rational or moral agents.

Given the somewhat conservative view adopted here, it is suggested in this article that while robots could certainly perform, at least in principle, some simple joint tasks (ones that are not strictly speaking joint *actions*, but are analogous to joint action), nevertheless, robots (or, at least, if it is insisted, the current generation, and near generations, of robots) are constitutively unable to perform morally significant joint actions since they cannot recognise moral properties, let alone act for the sake of moral considerations. Robots are not moral agents and nor are they, therefore, occupants of institutional roles properly speaking (although they may well be able to perform tasks, joint human/robot tasks, in institutional settings. Accordingly, they cannot be ascribed moral responsibility, whether individual or collective moral responsibility. Moreover, it is also suggested in this article that the program to find physical properties that robots are responsive to, and which reliably correlate with moral properties to any significant extent, faces very significant, arguably insurmountable, obstacles—obstacles which persist even if, as is the case, robots can identify and respond to new correlations, including ones which supersede old ones.

In tandem with the claim that robots are not moral agents, it is argued that robots are not rational agents in any strong sense and, therefore, cannot choose their ultimate ends, much less the ultimate collective ends of the institutions which make use of them. Moreover, since robots are neither rational nor moral agents they cannot exercise the reason-based moral judgments, including discretionary moral judgments of which many are means/end judgements (e.g., of the form, 'The end does (or does not) morally justify the means') required of institutional role occupants. Rather robots can only perform a restricted range of specialised tasks, including in institutional settings, that do not require moral judgments and do so in circumscribed domains under the tight control of human beings. Robots are, or ought to be, technological instruments under the control of human agents in the service of the collective goods definitive of the institutions in question, e.g., the collective good of security provided by police institutions (Ludwig, 2017; Miller, 2010; Searle, 1994).

In relation to the issue of human/robot joint action and (relatedly) robots (supposedly) occupying institutional roles alongside humans, it is important to get clear on the nature and importance of joint action and social institutions; hence my recourse to the above-mentioned collective end theory and the normative-teleological theory. Here we need initially to stress that a single human being can achieve relatively little acting alone, as opposed to acting as a member of a group, whether for good or for evil. Thus, the importance of cooperative action and, in particular, of joint action. Joint actions are actions in which a number of agents perform individual actions in the service of a common or shared end: a collective end in my parlance (Miller, 2001, 2007). Joint actions can be simple, brief actions involving a small number of agents, such as two men lifting a piano onto a truck. Or they can be complex, long-term actions involving large number of agents, such as building the Great Wall of China. Or they can be something in-between. Collective ends are ends which each contributing agent aims at in performing his or her contributory action, and which each agent possesses interdependently with the other agents doing so.

Complex, long-term, joint actions involving large numbers of agents typically require the agents to be organised. Indeed, in general, large groups cannot realise their joint ability (Miller, 2019b) to perform large tasks without being organised. Hence the importance of organizations and, in particular, institutions (Ludwig, 2017; Miller, 2010; Searle, 1994), such as universities, corporations, armies and organised occupational groups.

According to the normative-teleological theory favoured here, institutions are essentially joint enterprises (including joint epistemic enterprises) in the service of collective ends which are collective goods, e.g., universities have, or ought to have, as their collective end the collective good of

knowledge (broadly understood), the housing industry has, or ought to have, as its collective end the collective good of an adequate and sustainable supply of housing of reasonable quality and at a reasonable price, and so on.

It almost goes without saying that joint action, and the joint action of institutional role occupants in particular, is rendered far more efficient and effective by means of technology, including robotics. (Although this is not necessarily the case and, of course, efficiency and effectiveness are not the only considerations when it comes to evaluating joint actions—doubtless, for instance, Hitler's gas chambers were efficient and effective). Consider the use of robots in manufacturing plants. An important set of questions that arise at this point concern the extent to which robots could engage in joint action with themselves or with humans and ultimately, it might be suggested, occupy institutional roles thereby replacing human beings in those roles.

Regarding the latter point, we need to distinguish between institutions and organizations; and, more specifically, between institutional and organizational roles. It is important to note that on the stipulative definitions of organizations and institutions preferred here, and by contrast with institutions, organizations are, *qua organizations*, non-normative entities. So being an organization is not of itself something that is ethically or morally good or bad. This can be consistently held while maintaining that organizations are a pervasive and necessary feature of human life, being indispensable instruments for realizing collective ends. Collective ends are a species of individual end; but merely being an end is in itself neither morally good nor morally bad, any more than being an intention or a belief are per se morally good or morally bad. While this definition of an organization does not include any reference to a normative dimension most organizations do as a matter of contingent fact possess a normative dimension; most organizations are, therefore, also institutions. This normative dimension of organizations that are also institutions will be possessed (especially, though not exclusively) by virtue of the particular moral/immoral ends (goods) that such organizations serve, as well as by virtue of the particular moral (or immoral) activities that it undertakes.

Relatedly, note that the fact that humans, chimpanzees or robots can engage in joint or assisted actions (or, in case of robots, some analogue of intentional actions) does not in and of itself entail that they have normative status and, in particular, that they have moral status, for instance, that they recognise moral properties such as moral rights and duties, are able to empathise with others, and so on. Rather normativity is only *necessarily* involved in the minimal sense that these beings can succeed in performing actions or, in the case of robots, they can succeed or fail to do tasks as a car's brakes (once activated) can succeed in stopping its forward motion or fail to do so. Moreover, organisational action, understood simply as multi-layered structures of joint action (see below), is not, therefore, necessarily normative other than in this minimal sense, i.e., that organizational role occupants can succeed or fail in performing their tasks. This is in large part because collective *ends* are not necessarily collective *goods*, any more than individual *ends* are necessarily *goods*. That said, institutional roles, as opposed to mere organisational roles, such as that of teacher, doctor, police officer and engineer, are defined in part in terms of institutional rights some of which are moral rights and duties. These institutional rights and duties are derived in part from the collective goods that they serve and, since many of these collective goods are themselves moral goods, many of these derived institutional rights and duties are also moral rights and duties. Consider, for instance, a police officer's institutional and moral right to arrest an offender; right in part derived from the collective good of community security. Moreover, some of these institutional rights and duties are derived in part from prior constraining moral principles, such as the moral principle not to take innocent life—principle which constrains the institutional rights of police officers. In short, the individual and joint actions of institutional role occupants have a moral normative dimension.

Since our concern in this article is, firstly, with human/robot joint action and robots potentially occupying institutional roles and, secondly (and relatedly), with robots and moral properties, we need a serviceable account of joint action and of the more complex notion of multi-layered structures of joint action that underpins (in our favoured view) social institutions. We begin with joint action.

## Joint actions and social institutions

### Joint actions

Joint actions are (prima facie, at least) not the singular actions of collective entities, such as the supposed 'actions' of the nation-state, Australia. Nor are joint actions the aggregated singular actions of individual human beings, such as the set of actions consisting of Australians getting out of bed in the morning. Rather joint actions are actions involving a number of agents performing interdependent actions in order to realise some common goal. Examples of joint action are: two men lifting a piano onto a truck, two people dancing together, a number of tradesmen building a house and a group of robbers burgling a house.

There are various competing theories of joint action in the philosophical literature (Miller, 2001; Gilbert, 1992; Bratman, 2014; Ludwig, 2017). Here a strict individualist relationist account, namely the Collective End Theory elaborated elsewhere (Miller, 1992, 2001, 2007), is put forward.

Basically, CET is the theory that joint actions are actions directed to the realisation of a collective end. However, this notion of a collective end is a construction out of the prior notion of an individual end; accordingly, it is an attitude in what is referred to in the social ontology literature as the I-mode (by contrast with the so-called we-mode). Here the "I-mode" refers to a propositional attitude, e.g., an intention or belief, the content of which refers to a single individual, as in "I intend to go home". By contrast, the "we-mode" refers to a propositional attitude, e.g., an intention or belief, the content of which refers to a group, as in "We intend to win the war". Collectivist theorists, such as Searle (Searle, 1994) and Gilbert (Gilbert, 1992) maintain the we-mode attitudes are not reducible to I-mode attitudes. Roughly speaking, a collective end is an individual end more than one agent has, and which is such that, if it is realised, it is realised by all, or most, of the actions of the agents involved; the individual action of any given agent is only part of the means by which the end is realised. So a joint action simply consists of at least two individual actions directed to the realisation of a collective end.[1]

Accordingly, individual actions, x and y, performed by agents A and B (respectively) in situation s, constitute a joint action if and only if:

(1) A intentionally performs x in s and B intentionally performs, y, in s;
(2) A x-s in s if and only if (he believes) B has y-ed, is y-ing or will y in s and B y-s in s if and only if (he believes) A x-s or is x-ing or will x in s;
(3) A has end e, and A x-s in s in order to realise e and B has e, and B y-s in s in order to realise e;
(4) A and B each mutually truly believes that A has performed, is performing or will perform x in s and that B has performed, is performing or will perform y in s;
(5) Each agent mutually truly believes that (2) and (3).

In respect of this account the following points need to be noted.

First, with respect to clause (2), the conditionality of the action is internal to the agent in the sense that if the agent has performed a conditional action then the agent has performed the action in the *belief* that the condition obtains. Moreover, the conditionality of the agent's action is relative to the collective end. A x-s *only if* B y-s *relative to the collective end e*. So B's y-ing is not a necessary condition for A's x-ing, *tout court*. For example, it might be that A has some other individual end e1, and that A's x-ing realises e1 (in addition to e). If so, A will x, even if B does not y. But it remains true that A x-s only if B y-s *relative to the collective end e*. Again, A x-s *if* B y-s *relative to the collective end, e*. So B's y-ing is not a sufficient condition for A's x-ing, *tout court*. If for some reason A abandons the end e, then even if B y-s, A will not x.

Second, the notion of a collective end is an individualist notion. The realisation of the collective end is the bringing into existence of a state of affairs. Each agent has this state of affairs as an individual end. (It is also a state of affairs aimed at under more or less the same description by each agent.) So a collective end is a species of individual end. Thus CET is to be distinguished from the accounts of theorists such as John Searle (Searle, 1994) who favour irreducibly collectivist notions of collective intentions. Importantly, however, this individualism enables the account to be serviceable as a model for the doings of robots and, more specifically, of human/robot joint action. However, in using CET as a model for human/robot joint action, I am not, thereby, ascribing ends and intentions to robots (see below). Indeed, CET is serviceable as a model in part because it maintains an ontological segregation between human and robot, notwithstanding their cooperative interaction. By contrast, collectivist accounts of joint action postulate irreducibly collectivist mental states, e.g., irreducibly we-intentions and the like. These collectivist mental states either attach to the collective entity per se, i.e. the collective entity comprising human and robot (Gilbert, 1992), or imply that both robot and human are possessed of irreducibly we-intentions with respect to the joint action in which they are participants (Searle, 1994). However, it is unclear how such mental states can be possessed by a human/robot collective agent (whatever that might be) or that they could be possessed by robots (as well as their human co-participants in the joint actions in question), given humans and robots are evidently ontologically very different. Moreover, the CET account is doubly serviceable in that it does not entail any moral concepts and, therefore, is not in danger of begging the question regarding the possibility of robots being moral agents merely because they might be able to engage in joint action (on my account, as it turns out, in some attenuated sense of action). Further, this individualism and avoidance of moral concepts in the definition of joint action carries over to organisational action (as opposed to the morally encumbered notion of institutional action) since organisational action is to be understood in terms of layered structures of joint action (see below).

A collective end is the same as an ordinary individual end in that qua end it exists only in the heads of individual agents. But it is different from an ordinary individual end in a number of respects. For one thing, it is a *shared* end. By shared end I do *not* mean a set of individual ends that would be realised by qualitatively identical, but *numerically different,* multiple states of affairs. Rather I mean a set of individual ends that would be realised by *one single*

[1] The following definition is drawn from various previous publications of mine namely Miller 1992, Miller 2001 Ch. 2, Miller 2007 and Miller 2023a.

state of affairs; so the coming into existence of the one single state of affairs in question constitutes the realisation of *all* of the individual ends that exist in the heads of each of the agents. For another thing, although collective ends are shared (individual) ends, they are ends that are *necessarily* shared by virtue of being interdependent; they are not ends that are shared only as a matter of *contingent* fact. Suppose you and I are soldiers being shot at by a single armed drone. Suppose further that we both happen to see the drone at the same time and both fire at the drone in order to destroy it. I have an individual end, and you have an individual end. Moreover, my individual end is one that I share with you; for the destruction of the drone will not only realise my individual end, necessarily it will simultaneously realise your individual end. However, this is not yet a collective end; for the fact that I have as an end the destruction of the drone is not dependent on your having as an end the destruction of that very same drone, and vice-versa for you. The reason for this is that I can by acting alone destroy the drone; so my possession of that end (the destruction of the drone) is not dependent on your possession of that end. Likewise, your possession of that end (the destruction of the drone) is not dependent on my possession of that end.

The second related notion that I need to introduce is that of a multi-layered structure of joint actions. Organizational action typically consists in, what I have elsewhere termed, a *multi-layered structure of joint actions* (Miller, 2001, 2010). One relevant illustration of the notion of a layered structure of joint actions (in an institutional setting involving collective *goods*– so it is not *merely* organisational action) is a cybersecurity department comprised of three (let us assume for purposes of simplification) cyber teams: a cyber threat intelligence team (TI); an incident response team (IR), and an engineering team (EN). Suppose at an organizational level a number of joint actions ('actions') are severally necessary[2] and jointly sufficient to achieve some collective end, e.g., to prevent or mitigate malware attacks. Thus, the epistemic action of the TI team gives early warning to the IR team (which can act to prevent or, let us assume in this instance, mitigate a cyberattack) and, if necessary (as we assume it is in this instance), to the EN to enable it to 'patch' a defect in the system which the cyberattack is exploiting. Assume that the 'action' of TI is, in fact, a joint action, as is the 'action' of IR and the 'action' of EN. Moreover, assume also that the 'action' of TI, the 'action' of IR, and the 'action' of EN are severally necessary and jointly sufficient to achieve the collective end of preventing or mitigating the ongoing cyberattack, e.g., a virus; as such, these 'actions' taken together constitute a fourth joint action which is comprised of the three joint actions of TI, IR and ED (respectively).

At the first level there are individual actions directed to three distinct collective ends: the collective ends of (respectively) collecting and disseminating cyber threat intelligence, responding to the cyberattack, and removing the cyber system vulnerability. Thus, at this level there are three joint actions (of TI, IR, and ED, respectively). However, taken together these three joint actions constitute a single (second level) joint action. The collective end of this second level joint action is to mitigate the effects of the ongoing cyberattack; and from the perspective of this second level joint action, and its collective end, these (first level joint) constitutive actions are (second level) individual actions. Note that typically in organizations not just the nature, but also the quantum, of the individual contributions made to the collective end will differ from one team member to another.

Having elaborated our individualist analysis of the key notions of joint actions and multilayered structures of joint action, let us now elaborate our account of institutions and institutional role before turning directly to human/robot joint action and robot occupancy of institutional roles.

## The normative-teleological theory of social institutions

According to the normative-teleological theory (Miller, 2010), institutions are essentially joint enterprises in the service of collective ends which are, normatively speaking, collective goods. Thus, universities are joint enterprises in the service of the collective good of knowledge, while the housing industry is a joint enterprise in the service of providing an adequate supply of housing of reasonable quality and at a reasonable price (Miller, 2010 Ch. 10).

Clearly, different institutions exist to provide different collective goods. So there is a need to develop *special normative theories*, such as a normative theory of universities, a normative theory of the housing industry, a normative theory of policing, a normative theory of banks, and so on.

By the lights of the normative theory of institutions, technology, such as robotics and AI, ought to be inter alia a means to the collective ends of institutions. Thus, in the car industry, robots are a means to facilitate the efficient and effective production of cars. Again, in an army, robots are a means to facilitate the collective ends of an army, such as to afford protection to its personnel and of its civilians, as in fact happens in the case of robots that assist in the disarming of mines.

However, there is a further point to be made. According to the normative theory of institutions, our most important social institutions have, or ought to have, as their collective

[2] Here there is simplification for the sake of clarity. For what is said here is not strictly correct, at least in the case of many actions performed by members of organizations. Rather, typically some threshhold set of actions is necessary to achieve the end; moreover, the boundaries of this set are vague.

ends the production or maintenance of collective goods which consist in the satisfaction of aggregate needs, such as food, shelter, healthcare, security and education. Importantly, needs stand in contrast with desires or preferences. For unlike desires, needs are, firstly, objective, (e.g., food is objectively necessary), secondly, limited (e.g., a limited quantum of food is sufficient for one's needs) and, thirdly, generative of moral obligations since one is harmed if one's needs are not met.

Accordingly, *aggregate* needs in a community generate *collective moral responsibilities* to provide for those needs. Moreover, since the principal means, at least in modern societies, to discharge these collective moral responsibilities to provide food, security, education etc. are institutions, it follows that the institutional rights and duties of members of (respectively) agribusinesses, police services, universities etc. are derived from the collective goods that that these institutions exist, or ought to exist, to provide. Note that since these institutional rights and duties are derived from morally required (collective) goods, they are typically also *moral* rights and duties. Further, since technology is a necessary means by which institutions provide or maintain the collective goods in question, it follows that institutional actors have institutional rights and duties (which are also moral rights and duties) with respect to technology, e.g., a duty to provide staff with computers in universities and a duty on the part of staff to use them, a duty to provide army personnel with robots to disarm mines and a duty on the part of relevant personnel to use them to do so. Naturally, these derived moral rights and duties with respect to technology are rights and duties *other things being equal*, and other things might not be equal. Thus, an autonomous predator drone might be an efficient and effective means to kill enemy combatants but, nevertheless, morally impermissible by virtue of not having a human operator making the decision whether or not to pull the trigger (so to speak).

To reiterate: according to the normative-teleological theory, institutions are essentially joint enterprises that produce or maintain collective goods. Note the following additional points. Firstly, joint enterprises typically consist of multi-layered structures of joint action (see above) that typically utilise technology at each level.

> Secondly, joint enterprises or, at least those that constitute important institutions, typically consist of complex chains of joint activity (typically embedded in multi-layered structures of joint action). Consider a criminal justice institution. The criminal investigators, including forensic scientists, jointly collect and analyse the evidence, e.g., DNA samples, to determine whether a suspect is factually guilty or factually not guilty. If in the investigative stage the suspect is found likely to be factually guilty then the adjudicative stage commences. In large part on the basis of the evidence provided by the criminal investigators utilising technology in the prior investigative phase, the members of the jury jointly determine whether the suspect is legally guilty or legally not guilty.

Thirdly, institutions reproduce themselves by means of a structure of roles that are occupied by human beings who are, nevertheless, *replaceable* over time by other (appropriately educated or trained) human beings. Thus, a cohort of factory workers, police officers or academics at a given time is replaced by another cohort at some later date. Importantly, in relation to our concerns here, some of the tasks performed by the human occupants of roles can be performed by robots and perhaps, if suitably adjusted, *some* of these roles can be occupied by robots, i.e., the human occupants can be replaced by robot occupants. Such suitably adjusted roles would need to be ones which did not possess a constitutive moral normative aspect (see below). In both sorts of case there is typically human-robot joint action (in some sense of joint action). In the case of human occupants being replaced by robots the robots will presumably be participating in the joint actions formerly participated in by the human occupants they replaced. Let us now turn to a consideration of the core concept out of which institutions are analysed, at least according to the normative-teleological theory: joint action. More specifically, let us consider joint actions involving robots.

## Robots, joint action and institutional roles[3]

Consider the above-mentioned scenario in which two men jointly lift a piano onto a truck. The joint action consists of: A performing action x (A lifts one end of the piano) and B performing action y (B lifts the other end of piano); in x-ing, A has individual end, A(e), i.e., the end that the piano is on the truck; likewise, in y-ing, B also has e (B(e)) i.e., the end that the piano on the truck. There is both interdependence of action (x and y are interdependent) and of ends (A(e) and B(e) are interdependent) since neither A nor B can realise the end that the piano is on the truck by acting alone. (Therefore, this piano lifting scenario is unlike the destruction of the drone scenario mentioned above.) The collective end here is that the piano is on the truck, and it consists of a set of interdependent individual ends (A(e) and B(e)) each with the same content realised by the same state of affairs, namely, that the piano is on the truck.

Importantly, in our example, A could be a robot, as could B. Indeed, both A and B could be robots. Therefore, it might be argued, robots can perform joint actions, at least in some

[3] An earlier version of the material in this section appeared in Miller 2023b.

sense of joint action (call them 'joint doings') and, therefore, in some sense of action (call it a 'doing'). Note that the notion of full-blooded autonomy possessed by humans is out of place here in part because it implies moral autonomy and robots are not moral agents (see below) but in part also because robots do not have beliefs, intentions and other mental states with intentionality or 'directedness' (as opposed to functional or causal states) and such that the agent who possesses them could, at least in principle, be conscious of them. At any rate, let us refer to the tasks (understood as a neutral term with respect to mental states) performed by robots as *functional doings*.[4]

If A and B are both robots who can jointly place the piano on the truck then the following propositions seem to be true.

(i) Robot A does x and robot B does y and each does so in the service of a function, f, to place the piano on the truck; specifically, A does x in the service of function, f, B does y in the service of f, and Ax(f) and By(f) are interdependent;
(ii) A and B mutually adjust to one another's doings;
(iii) A is not merely assisting B (nor is B merely assisting A);
(iv) A is not under B's direction, nor is B under A's;
(v) A doing x and B doing y in the service of the function, f, is not merely a series of individual (functional) doings which do not serve a larger function. Such a series of individual doings (as opposed to joint doings) might consist of robot A carrying the piano to the truck having as its function to do so but *without having the piano on the truck as its further function*; robot B lifts piano onto the truck if the piano is sufficiently close to the side of the truck to do so and, in any case, irrespective of what A does, i.e. irrespective of whether A brings the piano to the side of the truck or it arrives by some other means;
(vi) Neither robot A nor robot B chose its function, i.e., to place the piano on the truck; rather this function was programmed into them.

Even if robots cannot perform joint action in the sense in which human perform joint actions, it seems that they can perform joint *doings* that replicate or mimick some basic joint actions that humans perform, such as the joint action of lifting a piano onto a truck. Moreover, at least in principle, the repetitive performance in recurring circumstances by robots of such basic joint doings in the service of a function might constitute occupancy of a simple organisational role in our favoured sense of an organisational role (as opposed to an institutional role), e.g., the role of jointly moving around and storing furniture in a warehouse. Or, at least, it might do so if these doings involved a degree of responsiveness to somewhat varying, albeit recurring, circumstances, and if these organizational roles were interdependent with other organizational roles in an organizational setting. In short, just as there is an analogy between the joint actions in the service of collective ends by human beings, on the one hand, and the joint doings in the service of functions by robots, on the other hand, so there is an analogy between the repetition of such joint actions constitutive of the occupancy of organisational roles by human beings and the repetition of joint doings constitutive of the occupancy of organisational roles by robots. Further, a human being and a robot can engage in a joint 'action/doing' in the sense of a joint pairing of an individual action in the service of the end of the human being with a doing in the service of the function of the robot where the state of affairs that realises the end is the numerically same state of affairs that realises the function. If such pairings (joint action/doings) were repetitive then they might constitute the exercise of human/robot interdependent organisational roles.

However, in occupying such a simple organisational role (whether single or interdependent with a human occupied role) the robots would not be choosing their functions; rather, as mentioned above, these would be programmed into them. Further, as we saw above, institutional roles (as opposed to mere organisational roles) are defined in part by deontic properties, notably rights and duties, many of which are also moral rights and duties. But robots (or, at least, the current generation of robots) are not moral agents capable of recognising, and acting for the sake of, moral considerations. Let us pursue further the issue of robots occupying institutional roles.

As discussed above, robots can adequately perform some simple organisational roles since they can perform individual and joint functional doings. Nevertheless, as mentioned above, there are reasons to think that robots cannot adequately perform many, if not all, institutional roles.

It might be argued that some institutional roles involving the performance of relatively complex tasks could, at least in theory, be occupied by robots possessed of sensors, vast data storage and retrieval capacity, and a prodigious calculative capacity. Moreover, it might be argued that a number or robots could not only perform simple joint tasks but also complex tasks. For at least in theory, a set of robots could replicate organisational action; each robot could occupy the organisational roles in a multi-layered structure of joint tasks. Consider three large sets of military robots linked to each other in each set in a manner that enables the members of each of the three sets to perform a given joint task (three joint tasks in total) but also each set to a single central

[4] Here I do not presuppose any particular account of functions other than they do not entail the existence of intentionality and, therefore, consciousness. Thus, functions in my sense might be given a causal analysis.

organising robot which coordinates and controls the performance of the three joint tasks of the subordinate sets of robots. However, there appear to two serious obstacles to the extension of these apparent possibilities to a wide range of institutional settings and roles. Firstly, robots cannot choose their 'ends' (i.e., their functions) including their collective 'ends' (i.e., their interdependent functions). The ultimate 'ends' i.e., functions of robots are programmed into them i.e. they are chosen by humans. The only functions 'chosen' by robots are ones that are the means to 'ends' i.e., functions that are chosen by humans. Secondly, robots cannot recognise and act for the sake of institutional rights and duties that are also moral rights and duties.

Regarding the first point, the ultimate 'ends' i.e. functions of robots are programmed into them i.e. they are chosen by humans. The only functions 'chosen' by robots are ones that are the means to 'ends' i.e. functions that are chosen by humans. So robots are mere instruments rather than rational agents (in the strong sense of agents that can choose their ultimate ends), even if they collect, store and use data and select certain 'doings' as means. More specifically, robots cannot participate (other than as instruments) in the process of determining institutional (collective) ends. Accordingly, they cannot occupy institutional roles that are involved in the process of determining institutional (collective) ends. Note that this process consists in large part in specifying the content of existing collective ends in the light of changing circumstances. Consider the institutional (collective) end of winning the war. What does winning, say, the Second World War consist of? Dropping atomic bombs on Hiroshima and Nagasaki? The latter decision was made very late in the day. Such collective ends are somewhat vague and stand in need of further specification in any given set of circumstances. However, circumstances change and thus the process of specification is continuous and subject to change. Moreover, the process of specifying institutional (collective) ends involves multiple actors at multiple institutional levels, e.g., the political leadership, including President Truman, who made the decision to win the war in this manner rather than some other manner, the military and intelligence leaders and their staff who planned the enterprise, including determining its likely consequences, the pilots and other air-force personnel who executed the bombing itself, the scientists and engineers who designed and built the atomic bomb etc. Moreover, each of these groups performing their respective joint epistemic actions and joint kinetic actions needed to be coordinated with one another and this coordinating activity itself consisted in part in joint epistemic actions. Further, this entire complex of activity and its ultimate collective end was inherently particular and involved a discretionary judgment (indeed, multiple discretionary judgments at different levels). In short, the design, construction and use of an atomic bomb was at the time unique, and unique in a manner that made its design, construction and use unable to be predicted at any significant level of specificity based on past bomb design, construction and use. Accordingly, this entire complex of joint activity and many of its constitutive joint action elements would have been immune to detailed prediction based on, for instance, machine-learning techniques. As such, at the very least many of the key institutional roles could not have been successfully occupied by robots dependent on AI. The more general point is that complex institutional activity typically has an inherent particularity that renders it immune to the substitution in key institutional roles by robots.

Regarding the second point which pertains to institutional moral rights and duties that are also moral rights and duties, robots do not care about anything or anyone, including themselves. Care in the sense involved here is a motivating *moral* attitude comprised of an integrated mix of cognitive, conative and affective elements, including moral beliefs, intentions and emotions. Moreover, arguably, beings without moral emotions have no capacity for genuinely moral judgements or actions. Thus, as Kant famously maintained, they might be able to comply with a moral rule not to kill by refraining from killing but they cannot comply for the sake of that moral rule understood as a moral rule; they are not acting out of the moral motive or moral sense. This brings us directly to the issue of robots and moral properties, including moral responsibility. Let us, then, elaborate further on this issue.

## Robots and moral properties

Unlike humans, robots do not have consciousness, cannot experience pain or pleasure and, as mentioned above, do not care about anyone or anything (including themselves). As also mentioned above, care is a motivating moral attitude comprised of an integrated mix of cognitive, conative and affective elements, including moral beliefs, intentions and emotions. But robots do not have moral beliefs and moral intentions, and certainly not emotions of any kind, including emotions comprised in part of motivating moral attitudes. Accordingly, a robot cannot literally be your friend or carer (in the sense of someone who literally cares about you (Sparrow, 2015; Wynsberghe, 2013), as opposed to merely providing assistance to you). Indeed, robots cannot recognize moral properties, such as moral rights and duties, moral virtues such as courage, moral innocence, moral responsibility, sympathy or justice. Therefore, they cannot act *for the sake of* moral ends or principles *understood as moral in character*, such as the principle of discrimination according to which *innocent* civilians should not be targeted in war. Clearly, robots are incapable of moral agency and, therefore, of being held morally responsible or of being possessed of

moral duties or obligations. Nor, arguably, do they have any morally relevant properties can could provide a basis for the ascription of moral rights as, for instance, the capacity for pleasure and pain might be held to provide a basis for the ascription of moral rights to animals.

In short, robots are not moral agents or (at least in their present mode of existence) moral beings in any non-trivial sense. Therefore, unlike humans, robots do not have inherent moral (as opposed to instrumental) value and, specifically, do not have value qua particulars, i.e., a single robot, unlike a single human being, is replaceable without moral loss.

As we saw above, the concept of institutional authority, as opposed to causal control, is fundamental to occupancy of institutional roles. Moreover, the concept of institutional authority entails institutional rights and duties. However, as we saw above, such rights and duties are frequently, if not invariably, dependent on *moral* right and duties (Miller, 2010, 2019a). It follows that robots cannot occupy institutional roles in any deep sense since they are immune to institutional authority. More generally, since robots are not moral agents then they cannot possess moral rights or, at the very least, cannot possess moral duties, and therefore cannot occupy a wide range (if not all) institutional roles since many of the institutional rights and duties constitutive of such roles are also moral rights and duties.

As a corollary, robots cannot be accountable institutional role occupants because they cannot be ascribed moral responsibility for their individual or joint doings. That is, they cannot morally justify their 'actions' and 'omissions' to an authority. Rather they can only be designed in such a way that they do their allotted tasks and don't cause unwarranted harm, or otherwise fail to do what is institutionally required of them. Accordingly, it is only the human operators, designers and owner of robots that can be held accountable for the doings of robots.

Notwithstanding the above deflationary account of robots, nevertheless, robots are evidently superior to human beings in various respects. Thus, so-called 'autonomous'[5] robots (hereafter, autonomous robots) are able to perform many tasks for more efficiently than humans, e.g., tasks performed in factory assembly lines, auto-pilots, driverless cars; moreover, they can perform tasks that are dangerous for humans to perform, e.g., defuse bombs (Miller, 2019a; Sparrow, 2007, 2021). In addition, autonomous robots can also be weaponised and in a manner such that robots control the discharging of their weapons against targets (and, possibly, the selection of their targets). Further, by virtue of developments in artificial intelligence, the robots have superior calculative and memory capacity. In addition, robots are quite literally without fear in battle; for, as mentioned above, they don't have emotions and care nothing for life over death, including their own destruction (which, of course, is not death in the sense applicable to humans).

New and emerging autonomous robotic weapons can replace some military roles performed by humans and enhance others. Consider, for example, the Samsung stationary robot which functions as a sentry in the demilitarized zone between North and South Korea. Once programmed and activated, it has the capability to track, identify and fire its machine guns at human targets without the further intervention of a human operator. Predator drones are used in Afghanistan and the tribal areas of Pakistan to kill suspected terrorists. While the ones currently in use are not autonomous weapons, they could be given this capability in which case, once programmed, and activated, they could track, identify, and destroy human and other targets without the further intervention of a human operator. Moreover, more advanced autonomous weapons systems, including robotic ones, are in the pipeline.

To reiterate, robots cannot recognize moral properties and cannot act for the sake of moral ends or principles *qua moral ends and principles*, such as the principle of discrimination applicable to combatants or the general moral principle to assist those in desperate need of food, water or shelter. Given the non-reducibility of moral concepts and properties to non-moral ones and, specifically, physical ones,[6] at best computerized robots can be programmed to comply with some *non-moral proxy* for moral requirements (Arkin, 2010; Miller, 2019a; Wynsberghe & Robbins, 2019). For example, 'Do not intentionally kill morally innocent human beings' might be rendered as 'Do not fire at bipeds if they are not carrying a weapon or they are not wearing a uniform of the following description'. However, this move faces significant indeed, arguably, insurmountable problems.

First, robots evidently require simulacra or proxies that reliably correlate with the possession of moral properties. Thus, they might require that the moral property of being an innocent civilian in some theatre of war reliably correlates with being a biped not in blue uniforms and not carrying a rifle. In the context of a conventional war in which combatants wear uniforms to clearly identify themselves as such in accordance with international law there may well be a high correlation between physical features to which robots are sensitive and moral properties, such as the innocence of civilians, to which they are not. However, similarly reliable correlations in the case of terrorists or spies may be hard to find, given terrorists and spies are in the business not

[5] Robots are not autonomous in the sense in which human beings are. For instance, human autonomy would seem to entail a capacity to choose one's ultimate ends and, arguably, to recognise moral properties.

[6] The physical properties in question would not only be detectable in the environment but also be able to be subjected to various formal processes of quantification and so on.

of identifying themselves as such but quite the opposite, namely, masking their identity. There is also the problem, at least in the case of spies and terrorists, that the numbers of spies or terrorists are simply too small for machine learning techniques, for instance, to discover any reliable correlations that would enable them to be readily identified. Moreover, in the case of terrorists and spies any systematic correlations that are discovered will not necessarily be reliable given that these correlations are based on past behaviour which is alterable at will. For instance, terrorists can change their clothing, avoid carrying weapons most of the time, and so on; likewise, spies can devise new modus operandi, change their outward behaviour, etc. and, importantly, do so in some instances based on their knowledge of the very past correlations that their protagonists are relying on.

Naturally, there are many reliable correlations, notably ones based on ongoing compliance with socio-moral norms, laws/regulations and conventions. Consider, for instance, the law/convention to drive on the left-hand side of the road (exploited by driverless cars) or the international law requiring combatants to wear uniforms which identify them as combatants (potentially exploitable by weaponised robots). However, these reliable correlations are typically known to all relevant parties; we do not require AI to enable us to identify them and, therefore, we do not need to rely on AI in order to make decisions based on our knowledge of them. On the other hand, to the extent that recognition of these norms, laws and conventions, and of compliance with them, requires the capacity for recognition of moral properties then AI, and machine learning in particular, is at a disadvantage. It requires non-moral proxies and they might not be available or, if available in some form, might not be reliable, including for the reasons already mentioned.

Second, as is well-known many institutional actors, such as combatants, police, politicians, doctors, lawyers, teachers, social workers, journalists, parents and spouses confront circumstances in which there are competing moral considerations and *discretionary moral judgements* need to be made, including in respect of whether the end morally justifies the means. In many of these circumstances there is no clear-cut rule to determine what is to be done or there are conflicting clear-cut rules or there is a clear-cut rule which might need to be ignored or a loophole found. Moreover, many of these circumstances are non-recurring or recur with significant differences. Such discretionary moral judgements rely on identifying, weighing and calibrating moral considerations and the nuances thereof. However, as has been made clear, robots cannot recognise moral properties as such, let alone engage in highly particularised discretionary moral judgments that resist formulaic (algorithmically expressible) treatment.

Third, even if, as is extremely improbable, a comprehensive set of reliable correlations could be found for, say, the principle of discrimination that governs combatants use of lethal force, that principle is conceptually interdependent with other principles (but not necessarily logically interdependent in a manner that enables the relation of interdependence to be algorithmically expressible). Thus, the principle of discrimination is conceptually interdependent with the principle of necessity and that of proportionality, and this interdependence would need somehow to be accommodated. Roughly speaking, the principle of discrimination forbids intentional targeting of innocent civilians[7] and, also, foreseeably and avoidably putting their lives at unnecessary risk. The latter clause conceptually implicates the principle of military necessity since a risk to civilians is unnecessary if the use of lethal military force which constitutes this risk is not militarily necessary. Therefore, the principles of military necessity and discrimination are conceptually interdependent. An analogous point can be made in relation to the interdependence of the principle of proportionality with the principle of necessity and the principle of discrimination. Accordingly, a conceptually independent, precise rule specification for the principle of discrimination would not work (nor, obviously, for the principles of necessity and proportionality). The general point to be made here is that moral principles are conceptually interdependent such that there no single moral principle correlates with some non-moral proxy.

Fourth, the precise rules in question would presumably be applicable to sharply defined, discrete, self-contained contexts involving the use of lethal force (analogous to the rules for driverless cars); otherwise, the robot would not be able to comply with them. However, in any such sharply-defined, discrete and self-contained context, be it a one-against-one lethal encounter, a firefight, an air strike or a battle, there will inevitably be moral considerations emanating from some other context (for example, another battle) or some larger context of which the discrete, self-contained context is an element (for example, the war as a whole), which bear upon it in a manner that morally overrides or qualifies compliance with the precise rule in question (or set of rules, for that matter[8]).

Fifth, and relatedly, moral principles, rights, duties and the like are not only interdependent, they are inherently general and applicable in verse diverse settings. As such, they

[7] Arguably, the component clause of the principle of discrimination, namely, the impermissibility of intentionally killing innocent civilians is logically independent of its second clause and of the other principles. This does not affect my argument. The principle of discrimination also applies to kind of weaponry uses. For example, biological weapons are indiscriminate.

[8] The sharply defined computerized rule conception could be complicated by adding meta-rules, for example. However, this would not make any material difference to the problems; it would simply elevate things to a higher level of complexity.

require significant adjustments from one context to another, i.e., their application calls for discretionary moral judgements (see above), as opposed to strict adherence to precise rules. Consider in this connection the moral principle of fairness required of teachers in relation to their students but also of police in relation to suspects; again, consider the moral principle of proportionality in respect of the use of lethal force required of police officers in peacetime as opposed to of combatants in a theatre of war. But robots do not have general intelligence of the kind required of rational agents, and of moral agents in particular. Hence robots are restricted to specialised tasks governed by precise rules to be applied in circumscribed domains.

## Conclusion

In summary, in this article I have discussed the idea that robots could perform joint actions, including morally significant joint actions, such as those performed by institutional role occupants. I have concluded that robots could perform, at least in principle, some simple joint tasks (ones that are not strictly speaking joint *actions*, but are analogous to joint action, i.e. joint doings in the service of functions), but that robots are constitutively unable to perform morally significant joint actions (or joint doings) since they cannot recognise moral properties as such, let alone act for the sake of moral considerations. Accordingly, they cannot occupy institutional roles that involve, as many do, the performance of morally significant actions, whether these be individual or joint actions (including as elements of multi-layered structures of joint actions). Moreover, I have also argued that the program to find physical properties that robots are responsive to, and which reliably correlate with moral properties to any significant extent, faces (arguably) insurmountable obstacles, obstacles which persist even if, as is presumably the case, robots can learn new correlations which supersede old ones.



## References

Arkin, R. (2010). The case for ethical autonomy in unmanned systems. *Journal of Applied Philosophy, 9*(4), 332–341.
Bratman, M. (2014). *Shared agency*. Oxford University Press.
Gilbert, M. (1992). *On social facts*. Princeton University Press.
Ludwig, K. (2017). *From individual to plural agency*. Oxford University Press.
Miller, S. (1992). Joint Action. *Philosophical Papers*, *21*(3), 275–299.
Miller, S. (2001). *Social action: A teleological account*. Cambridge University Press.
Miller, S. (2007). Joint action: The individual strikes back. In S. L. Tsohatzidis (Ed.), *Intentional acts and institutional facts. Essays on John Searle's Social Ontology* (pp. 73–92). Springer. https://doi.org/10.1007/978-1-4020-6104-2_3
Miller, S. (2010). *The moral foundations of social institutions*. Cambridge University Press.
Miller, S. (2019a). Machine learning, ethics and law. *Australasian Journal of Information Systems*. https://doi.org/10.3127/ajis.v23i0.1893
Miller, S. (2019b). Joint abilities, joint know-how and collective knowledge. *Social Epistemology., 34*(3), 197–212.
Miller, S. (2023a). Joint actions: We-mode and I-mode. In M. Garcia-Godinez & R. Melin (Eds.), *Tuomela on Sociality* (pp. 59–78). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-22626-7_4
Miller, S. (2023b). Robots, institutional roles and collective ends. In R. Hakli, P. Mäkelä, & J. Seibt (Eds.), *Social robots in social institutions* (pp. 16–22). IOS Press. https://doi.org/10.3233/FAIA220597
Searle, J. (1994). *The construction of social reality*. Penguin.
Sharkey, N. (2012). Killing made easy: From joysticks to politics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 111–128). MIT Press.
Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*. https://doi.org/10.1111/j.1468-5930.2007.00346.x
Sparrow, R. (2015). Robots in aged care: A dystopian future? *AI and Society Published Online First*. https://doi.org/10.1007/s00146-015-0625-4
Sparrow, R. (2021). Why machines cannot be moral. *AI & Society: Journal of Knowledge, Culture and Communication*. https://doi.org/10.1007/s00146-020-01132-6
Wynsberghe, A. (2013). Designing robots for care. *Science and Engineering Ethics, 2*, 407–433.
Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*, 719–735.