sciendo

# Employing Robots

**Carl David Mildenberger**
University of St. Gallen

DOI: 10.2478/disp-2019-0013 BIBLID [0873-626X (2019) 53; pp.89–110]

**Abstract**
In this paper, I am concerned with what automation—widely considered to be the "future of work"—holds for the artificially intelligent agents we aim to employ. My guiding question is whether it is normatively problematic to employ artificially intelligent agents like, for example, autonomous robots as workers. The answer I propose is the following. There is nothing inherently normatively problematic about employing autonomous robots as workers. Still, we must not put them to perform just any work, if we want to avoid blame. This might not sound like much of a limitation. Interestingly, however, we can argue for this claim based on metaphysically and normatively parsimonious grounds. Namely, all I rely on when arguing for my claim is that the robots we aim to employ exhibit a kind of autonomy.

**Keywords**
Artificially intelligent agents, robots, autonomy, alienation, role-responsibility.

## 1 Introduction

A future in which artificially intelligent agents perform the majority of those jobs nowadays done by humans has been pictured both as a utopia and a dystopia. In this paper, I am not concerned with what such a future holds for us—but what it holds for the artificially intelligent agents we aim to employ. My guiding question is whether it is normatively problematic to employ artificially intelligent agents like, for example, autonomous robots as workers, given what we do to them if we do so. The answer I propose is the following. There is nothing inherently normatively problematic about employing autonomous robots as workers. Still, we must not put them to perform just any work, if we want to avoid blame.

That might not sound like much of a limitation. Interestingly, however, we can argue for this claim based on metaphysically and normatively parsimonious grounds. Namely, I will neither assume nor defend that robots are conscious or self-aware, that they are able to experience pleasure and pain, that they possess human dignity or personhood, or that they have rights. This is because all of these ideas are highly controversial (e.g. Searle 1980, Dennett 1997, Asaro 2011, Gunkel 2018). By contrast, all I rely on when arguing for my claim is the idea that what we want to do is to employ artificially intelligent agents which exhibit a kind of autonomy as workers.

A direct consequence of this low metaphysical and normative involvement is that I am dealing with a real problem here. Artificially intelligent agents already are, and increasingly will be, autonomous agents of some form—even if they might never be conscious or sentient. Consequently, what I am arguing against is the position that we will only encounter normative problems with employing robots *later*; for instance, should we choose not to keep them slaves (cf. Bryson 2010). If we do not want to breed and employ a new but still somehow deprived working class, then there are reasons to worry *now*—and not only once we agree that we actually have succeeded in building sentient and conscious beings for the purpose of letting them work for us.

I shall first outline in which sense I take the robots we aim to employ to be autonomous (Section 2). Notably, I propose a technology-inspired definition of *thin autonomy*. I then discuss what normatively follows from us employing thinly autonomous robots, namely a certain *role-responsibility* on our part (Section 3). We are role-responsible for not putting the thinly autonomous robots to perform *alienating* labor. In Section 4, I take an agnostic turn. Even if there is a cogent argument that it would be blameworthy to let thinly autonomous robots perform alienating labor, in the end we lack a solid epistemic foundation for judging whether we (already) do so.

## 2 Automation and autonomy

At the time of writing this, some artificially intelligent agents already exhibit a kind of autonomy. Let us call it *thin autonomy*. To understand what I mean by thin autonomy, we have to look to tech-

nical contexts. Consider the example of *Google*'s Go-playing software *AlphaGo*.

In 2016, *AlphaGo* was the first Go-software to beat a professional human player in a series of games (Google 2016). *AlphaGo* acquires its Go-playing skills based on machine learning; specifically, by building up an artificial neural network (a deep learning method based on reinforcement learning) through extensive training. The version of the software dubbed *AlphaGo Zero* does so purely relying on games it played against itself, after having been told nothing but the basic rules of Go. In 2017, *AlphaGo Zero* surpassed the playing strength of all prior versions of *AlphaGo* within 40 days, purely based on self-play, without relying on any human intervention or historical data (Silver et al. 2017).

> As [AlphaGo Zero] plays, the neural network is tuned and updated to predict moves …. This updated neural network is then recombined with the search algorithm to create a new, stronger version of AlphaGo Zero, and the process begins again. In each iteration, the performance of the system improves by a small amount, and the quality of the self-play games increases … This technique is more powerful than previous versions of AlphaGo because it is no longer constrained by the limits of human knowledge. … AlphaGo Zero also discovered new knowledge, developing unconventional strategies and creative new moves. (Deepmind 2017)

It is such self-learning mechanisms that lie at the heart of thin autonomy. Being given only a minimum amount of input, the AI learns itself to perform a certain activity with a given final goal, eventually coming up with new ways of doing things that have not been foreseen by humans. The AI performs instrumental calculations (considering moves, alternative tactics, etc.) and develops its own instrumental goals (e.g. to achieve a favorable position on the board) based on which it then executes moves.

Importantly, an artificially intelligent agent does not have to be a *strong AI* (Searle 1980) to be thinly autonomous. We may define the goal of creating a strong AI to be to "create artificial persons: machines that have all the mental powers we have" (Bringsjord and Govindarajulu 2018). If we actually succeeded in building such an artificially intelligent agent, then pretty much by definition this agent would be *fully autonomous*. An agent can be said to be fully

autonomous—i.e. autonomous in the classical philosophical sense of the term (as typically applied to human beings)—if it has "the capacity to be [its] own person, to live [its] life according to reasons and motives that are taken as [its] own and not the product of manipulative or distorting external forces" (Christman 2018). For instance, a fully autonomous artificial agent would be able to authentically embrace a final purpose like, say, pacifism. (Imagine a lethal autonomous weapon system as autonomous as being capable of becoming pacifistic.) By contrast, all that is required for an artificially intelligent agent to be thinly automous is that it is capable of self-learning, in the sense of performing its own instrumental calculations and its own instrumental goal-setting in view of an externally determined final goal.

One indicator that we are dealing with a thinly autonomous artificial agent is the unpredictability of its next move. Even an experienced Go-player with a thorough understanding of what a certain position on the board calls for will often be surprised by *AlphaGo Zero*'s next move.[1] To be sure, we are not in a state of complete unpredictability. One way in which we can predict what thinly autonomous artificial agents are going to do is what Bostrom calls *predictability through design*.

> If … the designers of … [a thinly autonomous] agent engineer the goal system of the agent so that it stably pursues a particular goal set by the programmers, then one prediction we can make is that the agent will pursue that goal. (Bostrom 2014: 108)

We can stably predict that the final goal of *AlphaGo Zero* will be to win a game of Go, and that its actions will be in line with this goal. More generally speaking, with the thinly autonomous agent's final goal being transparent for us (and determined by us), only the ways in which it will try to reach this final goal are unpredictable. It is characteristic of thinly autonomous agents that they are only autonomous with respect to their choice of means and instrumental goals, but not with respect to their final goal.

And we can even make some reasonable predictions about the

[1] Unpredictability cannot be more than an indicator. This is because, for instance, a complex machine which chooses its next action randomly (without learning anything from its actions) might be equally unpredictable.

instrumental goals and means most thinly autonomous agents will employ to reach their final goals. As there are some means that are useful to reach almost any goal, there is *predictability through convergent instrumental reasons* (Bostrom 2014: 108). Examples of such predictable instrumental goals are self-preservation (measures taken by the robot to ensure its survival), cognitive enhancement ("improvements in rationality and intelligence will … improve … decision-making, rendering the agent more likely to achieve its final goals" (Bostrom 2014: 111)), and resource acquisition.

Still, unforeseen "moments of creativity" remain a live possibility and testify to the artificial agent's thin autonomy. On the one hand, these can be actions which, in principle, would have been predictable—only that nobody thought very hard about them. One example is the way in which the artificial intelligence *OpenAI* played the multiplayer video game *Dota 2*. In *Dota 2* each player controls an avatar. The goal is to defeat the opposing team's avatars in player versus player combat and to destroy the opposing team's base. When *OpenAI* played *Dota 2*, it proceeded to do things which puzzled its designers. It seemed to randomly and unnecessarily take damage from software-controlled avatars early in the game. At first, the designers thought they had made a programming mistake. "We thought perhaps we needed to roll back, but noticed further gameplay was amazing, and the … [early] behavior was *baiting* the other … [avatars] to be aggressive towards it" (OpenAI 2017, my emphasis). That is to say, *OpenAI* came up with the idea that, in order to make the other avatars attack it, it needed to look weak, when in fact it had an advantage.

The creativity we witness here with this self-learned baiting strategy is just an instance of thinking outside the box. With the thinly autonomous agent possessing "fresh eyes" and a non-human mind, it might act in ways that are not entirely opaque to us—but that we just had not considered.

On the other end of the spectrum are cases like *AlphaGo Zero* coming up with fundamentally new moves we can meaningfully call creative and unpredictable, as they are moves human beings have not been able to come up with in the last 2500 years.

Thinly autonomous artificial agents which can fulfill tasks more complex and more varied than playing games in stunningly good ways have yet to be developed. But little imagination is needed to

see that thinly autonomous agents advanced enough to revolutionize, say, economic production processes are a technological possibility. Thin autonomy will be a sought-after feature of artificial agents used in production in the future. First, because it makes them more resilient to unexpected events and thus less prone to breakdown. Second, because by granting them thin autonomy, they might be able to identify better ways to perform the task in question.

## 3 Role-responsibility and alienation

I do not want to argue that there is a *direct* normative consequence of artificially intelligent agents' thin autonomy. Notably, I do not want to argue it is intrinsically valuable, giving us a *pro tanto* reason not to violate it.[2] What I want to argue instead is that there is an *indirect* normative consequence of artificially intelligent agents' thin autonomy. Namely, if we employ an artificial agent like, say, a robot, and if this robot is thinly autonomous, then we take on a form of *role-responsibility* (Hart 1968: 212–3). Let me first briefly elaborate on the concept of role-responsibility. I will then argue why we take it on if we employ a thinly autonomous robot—and also outline the normative consequences this has for us.[3]

### *Role-responsibility*

Hart describes role-responsibility as the idea that, "whenever a person occupies a distinctive place or office, … he is properly said to be responsible for the performance of the duties [attached to this role], or for doing what is necessary to fulfill them" (1968: 212–3). Classic

[2] One point which would speak against such an argument is a classic Kantian one. Namely, that only beings which are more fully autonomous are part of the moral community, and thus only violating such beings' fuller autonomy would be normatively problematic (Fox 2007).

[3] For ease of expression, I shall henceforth use "robot" interchangeably with "artificially intelligent agent". Robot is derived from Czech "robota", meaning forced labor, with Čapek (1924) being the first to introduce the term to refer to soulless automata performing tasks. Thus, this term nicely captures what we are concerned with here: letting artificially intelligent agents perform tasks in a work context.

examples are a captain's responsibility for his ship, or a judge's responsibility to give instructions to the jury before they begin deliberating. Several points are noteworthy about the concept in our context.

First, although role-responsibility may sometimes be connected to us being responsible for the actions of that entity we are role-responsible for (e.g. in the case of parents' role-responsibility for their children), this is not necessarily the case. As a judge, for instance, we are not responsible for the decision of the jury—only that the decision be reached in due process. This means that in case we are role-responsible for thinly autonomous robots, this does not imply that we are responsible for their actions.

Second, what we are responsible for is, in the widest sense, the "well-being" of the entity we are role-responsible for (Hart 1968: 213). As a captain we have to care for our crew and the ship. A judge takes care, as it were, of the "well-being" of the legal system by ensuring due process. This focus on caring for the well-being of that entity we are role-responsible for can plausibly be derived from the power we have over that entity due to our office or role. We basically act as trustees who must not abuse our power. Therefore, in case we are role-responsible for thinly autonomous robots, we are likewise responsible for their "well-being".

Third, being role-responsible for non-human beings or non-living entities is quite a natural thing. As a committee secretary, I might be responsible for producing correct meeting minutes (Waller 1993). And the role-responsibility of a captain for his ship might quite literally also be understood as a responsibility for its physical integrity. This also means that the entity we are role-responsible for does not need to have moral status (cf. Kamm 2007: chap. 7). Consequently, role-responsibility for thinly autonomous robots is not a non-starter, in the sense that they would clearly be the wrong object for that kind of responsibility.

Fourth, Waller (1993, 2011: chap. 6) emphasizes how role-responsibility is a kind of "take-charge-responsibility" (1993: 46). It differs from moral responsibility notably in how we acquire it.

> I can certainly take responsibility—take-charge-responsibility—for the role of committee secretary (I can volunteer for the job, or just take on the task when no one else does it); but I cannot simply take … [moral] responsibility. … My wish to take moral responsibility for …

> [a certain] failure may be touching, but it will carry no weight in actually establishing [moral responsibility]. (Waller 1993: 47)

In much the same way, I can dismiss take-charge-responsibility rather easily—by stopping to fulfill a certain role—but I cannot simply dismiss moral responsibility. That is to say, we can take on and dismiss role-responsibility (as an instance of take-charge responsibility) in a more self-determined way. This is important to keep in mind when thinking about the extent to which we are role-responsible for robots. Role-responsibility lends itself to capturing the idea that not all of us are always and necessarily responsible for all of them.

Finally, taking on role-responsibility by taking on a specific role "does not in itself … commit one to any favourable or unfavourable assessment" (Haydon 1978: 47). Typically, we are not to be praised or blamed simply because we are judges or committee secretaries.[4] This means that taking on the role of an employer (who also chooses to employ robots) is not in itself praise- or blameworthy. However, taking on role-responsibility "does leave scope for assessment of the way in which … [we] discharge it" (1978: 47). This means that although role-responsibility as a kind of take-charge-responsibility does not directly ground moral responsibility, it may do so indirectly. If I take on a role, but discharge it in a bad way, I might be properly blamed for this.

## *Why we are role-responsible for robots*

If we take on the role of an employer, then we *ceteris paribus* are role-responsible as an employer.[5] This is because employers are in a position of power as regards those whom they employ; notably, employers may order their employees to do this or that. Anderson (2017: 134–6) provides us with some examples for employers abusing this power. For instance, an employer may use her power to get her employees to wear diapers at work (so that they need fewer bath-

[4] Honorary offices (on the positive side) or being a "Blockwart" in Nazi Germany (on the negative side) seem to be examples for exceptions to this rule.

[5] When I speak of "employer", "employee", and "employment" in the following, I do not mean to use these terms in the narrow sense, i.e. as closely associated with wage labor, labor contracts etc.

room breaks), or to tolerate sexual harassment as a "natural" aspect of working in a restaurant. An employer's role-responsibility clearly speaks against such practices.

With respect to establishing that employers are role-responsible not only for their human employees, but also for the robots they employ, it is important that there is one common aspect in both types of employment relationship. Namely, the aspect that the employer has the power to control the employee's/robot's options and actions to further her own purposes (like raising productivity or maximizing profit). In employment relationships, the minds and bodies of those employed, so to speak, become the employer's tool.

We might not take this to be a big normative issue. After all, an employee voluntarily chooses to work for a certain employer and gets compensated for his work.[6] But relying on Hayek's discussion of coercion, we can at least speak of a *coercive element* that is present "when one man's actions are made to serve another man's will, not for his own but for the other's purpose, … [when his] mind is made someone else's tool" (1960: 133). What is characteristic of relationships featuring this coercive element is that the coercee still has a certain limited power to choose his course of action. But this happens in an environment where

> the alternatives are determined … by the coercer so that … [the coercee] will choose what the coercer wants. He is not altogether deprived of the use of his capacities; but he is deprived of the possibility of using his knowledge for his own aims. (Hayek 1960: 134)

Employment relationships are a prime example of relationships featuring this coercive element. And to the extent that the employee serves the employer's will, not for his own but for the employer's purpose, they are normatively worrisome. To be sure, Hayek (1960: 138–9) stresses that there are good reasons to accept this practice *all things considered*, and that employment as such does not amount to full-blown coercion. Still, the coercive element such relationships feature is *pro tanto* normatively problematic. And it is the employer's coercive power in this respect which, so to speak activates her role-

[6] However, compare Zimmerman (1981) and Stevens (1988) on coercive job offers and wages.

responsibility; as a mechanism to check this power.

It is the robot's thin autonomy which underlies the idea that there are analogous normative worries in the relationship between employer and robot. Unless we are dealing with a self-learning robot able to practice its own instrumental calculations and to come up with its own instrumental goals, it makes no sense to fear (or even speak of) making the robot subservient to our will in Hayek's sense. We rightfully do not worry about coercive elements in our working relationship to a hammer, or even to a complex machine. But we should worry when we employ self-learning robots which are sufficiently autonomous in their instrumental reasoning and goal-setting, so that we can meaningfully speak of us having the power to make them subservient to our will.

To be sure, we do not need to *threaten* robots to make them subservient to our will (Hayek 1969: 138–9). At the same time, I think it would seem weird to completely count robots' docility against them; in the sense of exculpating the employer from the very start. As if an employer were not role-responsible for an employee who subserviently caters to the employer's every whim from the very beginning of the relationship. Also note that common ways to dispel doubts about potential coercion—namely that employees voluntarily choose to work for the employer and get compensated for their work—are absent in the relationship between robot and employer.

Importantly, whether we have a *pro tanto* reason not to violate robots' thin autonomy because we are role-responsible towards them depends on the role we take on with regard to them. If we take on the role of employer and choose to employ them to use their "minds" as tools for our purposes, this is different from when we take on the role of users or designers of robots. Consider, for instance, Bostrom's rogue AI case. "An AI, designed to manage production in a factory is given the final goal of maximizing the manufacture of paperclips, and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips" (2014: 123). For the AI's designer, there is *no* reason not to violate the AI's thin autonomy. The designer even has a very good reason to interfere because of his specific role-responsibility. Consider that while professional codes of conduct for managers or employers typically emphasize how it is their ethical duty to empower and lead their employees, codes

of conduct for designers and engineers emphasize different aspects: precautionary principles, risk assessments, responsibilities for safety, or to further a society's overall welfare (e.g. Martin and Schinzinger 2004: chap. 10, van de Poel and Royakkers 2011: chap. 8). The role-responsibility of a designer (with the task of designing a certain kind of thinly autonomous agent) is not identical to that of an employer (who wants to maximize her gains via employing such an agent). For the latter, being in control over the robot leads to a power which needs to be checked by her role-responsibility. For the former, being in control over his product is the very thing role-responsibility requires.[7]

To conclude, we are not by default morally responsible for what we do to thinly autonomous robots. It is not their thin autonomy as such which has direct normative consequences for us. But their thin autonomy renders relationships featuring coercive elements possible. If then somebody makes the decision to take on the take-charge role-responsibility of an employer, this generates normative consequences for her. We *make ourselves subject* to judgments of praise and blame with respect to how we discharge of the role as employer of robots.

## *What we are role-responsible for*

It makes intuitive sense that what we are role-responsible for as employers at least partly depends on those whom we employ. Take the example of slavery. It is inacceptable that an employer owns those human beings she employs. Yet, nowadays robots naturally are considered somebody's property. And it has been argued that they necessarily should be (Bryson 2010). Bryson holds that to further humanize robots—for example by trying to tap our full technological potential in order to build robots which are legitimate bearers of rights—would dehumanize real people. Now, if robots should be somebody's property, then this is a big obstacle for the idea that an

[7] Arguably, it even is part of a designer's role-responsibility to prevent an artificially intelligent agent from committing suicide; i.e. not only to prevent damage the AI does to human beings, but to itself. This might the case, for instance, if there are two AIs with the same goals, so that the less potent decides to destroy itself in order to allow the more potent to proceed more efficiently (Omohundro 2008).

employer's role-responsibility for her robots implies non-slavery (as it does for her human employees).[8]

Or consider exploitation. Human employees might be said to be exploited if their wages are too low; e.g. if they suffice for maintaining their labor power, but do not grant them any share of the surplus value created by their labor (Marx [1867] 1992). But robots do not receive wages in the first place—and plausibly so. Thus, while a responsibility not to exploit one's human employees makes sense, it is even conceptually difficult to get the idea off the ground for robots.

The examples of slavery and exploitation already show that we seem to be role-responsible for comparably little when it comes to employing robots. This makes sense. Intuitively, human beings deserve more protection than robots; especially if we do not assume them to have rights, be persons, experience pain, etc. Still, I also think that it is a part of the role-responsibility of an employer not to *alienate* those she employs—and that this applies to both human employees and robots.

Consider the example of George. George works for a producer of weapons of mass destruction. Suppose that for George there are utilitarian reasons to do so, even if he himself is a pacifist, and that these reasons are why he performs this job (cf. Williams and Smart 1973: 97–9). George's job features the coercive element described by Hayek. In fact, George's boss not only uses George's body and mind *not for George's but for her own* purposes (to produce more weapons), but she uses them directly *against* George's (pacifist) purposes.

This is an instance of alienating labor for George. As the employer uses George's mind against his purposes, a widely accepted condition of many theories of alienation is met; namely, that a subject (George) and an object (his mind) get separated in a problematic way (cf. Leopold 2018). More specifically, and relying on Jaeggi's (2014)

[8] Gunkel (2014, 2018), for one, argues that robots not only potentially *can* but *should* have rights, and should not be considered property, thus contradicting Bryson. However, Gunkel's argument is rather radical, as it is "questioning the systemic limitations of moral reasoning, requiring … a thorough examination of the way moral standing has been configured in the first place" (2014: 113). I do not want to delve further into this debate, as it is somewhat opposed to this paper's spirit of looking for a less normatively demanding basis for granting robots a certain moral status.

recent theory of alienation, we may say that the core of alienation is a "relation of relationlessness" (2014: 1), i.e. a condition marked not by the absence of a relation from a certain subject to a particular object, but by a deficient relation, a lack of proper connection. Because his employer uses George's mind against his purposes, George is unable to make his labor his own, to "appropriate" it. George's case is an instance of alienating labor as the labor is anti-emancipatory in character, in the sense that his employer prevents George's self-realization (Jaeggi 2014: 32ff.).

What an employer's role-responsibility does, generally speaking, is to block the employer from alienating her employees. If the employer not only uses her employee's mind for her own purposes, but does so in an alienating way, this is incompatible with her having to take care of the well-being of her employee. For this would mean that the employer basically *turns the employee's mind against himself*.

Now, why is alienating labor—in contrast to slavery and exploitation—part of an employer's role-responsibility not only for her human employees but also for the robots she chooses to employ? This is because of a certain connection between alienation and autonomy. Namely, there is a significant risk that we push those we employ to perform alienating labor precisely when we violate their autonomy in the context of work, e.g. by telling them what to do and in which way. A being's autonomy not only "activates" the employer's role-responsibility in general. It is also what grounds worries of alienation in particular, i.e. that we may be hindering a being's self-realization. Put differently, if unjust property relationships are a prerequisite for problematic slavery, and if unfairly low wages are a prerequisite for exploitation, then a certain autonomy is the prerequisite for alienation with its anti-emancipatory core. Because of its emphasis on process and form rather than content and a being's essence, Jaeggi's theory of alienation is particularly suitable for capturing this point (cf. Jaeggi 2014: chap. 2). It allows us to see how interfering both with human employees' and thinly autonomous robots' instrumental reasoning and goal-setting in a work context can be alienating.

Still, there are some important differences between human employees and robots which need to be addressed. First, note that being alienated does not necessarily presuppose the subjective feeling of being alienated on part of those who are alienated. There is not only

*subjective* alienation, but also *objective* alienation (Leopold 2018). One classic example of an instance of objective alienation is the passage where Marx suggests that even capitalists are alienated in a capitalist system. Because of the economic structures and institutions of capitalism, neither proletarians nor capitalists get to practice self-realizing activities in capitalism—but the latter at least do not feel devastated in light of their being alienated, but "at ease" ([1845] 1975: 36). Thus, it is not the case that robots, simply because we assume them to be unable to experience being alienated, could not be alienated in the first place.[9]

Second, beings need not be fully autonomous to be alienated. The case of George might be interpreted to suggest that employers being role-responsible for their robots is a hypothetical scenario. After all, we might never achieve to build a strong AI able to authentically embrace purposes like, say, pacifism. But this interpretation would be misguided. To be sure, if a being is fully autonomous, then examples for how employers may use their power in an anti-emancipatory way are particularly easy to come up with. You just highlight how the employer may override her employee's authentically embraced purposes, thus hindering his self-realization. But full autonomy on the employee's part is not a necessary prerequisite for alienating labor.

Consider Geoff. Geoff, a weapon enthusiast, likewise works for the producer of weapons of mass destruction. He identifies with his employer's goal to produce and sell ever more weapons. Unfortunately, his employer holds a personal grudge against him. As a consequence, she does not allow him to do some production-related calculations efficiently, i.e. by using the sophisticated mathematical methods that Geoff masters. Instead, she successfully urges him to fulfill calculation-heavy tasks by using an abacus; and sometimes even by counting beans.

Even if Geoff's boss only interferes with Geoff's instrumental reasoning and goals, rather than with his bellicose purposes, this is an instance of her using his mind as a tool, not for his, but for her and against his purposes—and thus of alienating labor. Her purpose might no longer be to increase production or maximize profit.

[9] Objective alienation is compatible with Jaeggi's account—even if Jaeggi herself does not explicitly develop this line of thought (cf. 2014: xv).

Rather, her purpose might be to humble Geoff. But clearly, this is a way of hindering Geoff's self-realization and of using her power in anti-emancipatory ways.

What Geoff's case shows is that the employee does not need to have any authentically embraced final goals or purposes for there to be alienation. The normative worries start once the employer has the possibility to violate the employee's (lower level) autonomy with respect to instrumental reasoning and goal-setting. This is because as soon as the employer has the possibility to override (or actively dull) the employee's mind with respect to instrumental reasoning and goal-setting, we can meaningfully speak of alienating labor. There might still be a difference in degree here. It seems that we may alienate a certain being more "effectively", in the sense of alienating it faster and in a more thoroughgoing way, by preventing its self-realization with respect to final goals and purposes rather than with respect to instrumental goals. But both kinds of anti-emancipatory action qualify as making a being perform alienating labor.[10]

To conclude, because (i) non-alienation is a widely accepted candidate for being part of an employer's role-responsibility, and because (ii) there is a link between violating a being's autonomy and alienation, and because (iii) beings need not be fully autonomous to be alienated, and because (iv) alienation does not require (subjective) feelings or consciousness on the part of the alienated being, it is *prima facie* plausible that we should be role-responsible for not alienating the robots we employ.

[10] One might object that what is required to meaningfully speak of putting a being to perform alienating labor is that that being has its "own" goals, in the sense of goals it positively values. If this were the case, then one might again wonder whether artificially intelligent agents will ever be able to do this. I agree that if we want to meaningfully speak of putting a being to perform alienating labor by interfering with its (instrumental) goals, it is required that those goals are the being's "own" goals in some sense. But I also think that it suffices that they are self-chosen with respect to the self-learning *process* that lead to their adoption (say some kind of deep learning method). Put differently, valuation is not required for making a goal a robot's goal. Once again, consider Geoff. For his employer being able to alienate him, it is not required that Geoff is particularly invested in the mathematical method he wants to use to perform the calculations. What matters is the employer's anti-emancipatory interference.

## 4 Alienation and opacity

It should be clearer by now why I think that we must not put robots to perform just any work, if we want to avoid blame. The reason why I think this claim to be true is that we are role-responsible for not putting thinly autonomous robots to perform alienating labor.

Yet, this claim clashes with something I consider a widespread intuition; namely, the intuition that we may put robots to perform just any work. *Prima facie*, it seems like the only work we must not put robots to is immoral work. In this sense, the claim that we must not put robots to perform just any work seems trivially true. If the mafia had a hitman robot, it seems it would be wrong to put it to carry out hit jobs. That is because prompting hits jobs is generally wrong, no matter who does the killing.[11] Now, the discussion about which kinds of work are intrinsically immoral (and thus should be done by nobody) should be kept separate from the discussion about which kinds of work we must not put thinly autonomous robots to. Yet, other than immoral work, it seems like there are no moral limits for the kind of work we may put robots to. For instance, we regularly employ them to perform "bad" or extremely dangerous work, like defusing bombs or working in toxic environments. If anything, putting robots to perform kinds of work that human beings shun (e.g. for moral reasons) seems to be a particularly meaningful use of robots.

However, I think the widespread intuition that we may put thinly autonomous robots to perform just any work lacks a solid epistemic foundation. An agnostic stance is called for. For we cannot reliably judge whether and when we alienate robots.

The reason for this is thinly autonomous robots' *opacity*. We lack a reliable epistemic access to a thinly autonomous robot's internal instrumental calculations and goal-setting. Especially as regards self-learning robots which rely on deep learning methods, this is a widely known issue in AI. It is a characteristic of deep learning-methods that even the designers do not fully know why and how exactly the thinly autonomous agent succeeds in solving the given task—or what

[11] A special case is the use of autonomous robots for warfare purposes, i.e. the case of lethal autonomous weapon systems (e.g. Sparrow 2007). This is a highly specialized discussion I do not want to go into. I am concerned with more mundane cases of work here.

precisely went wrong in case of failure (e.g. Pearl 2018). Robots' opacity makes it exceedingly difficult to tell what kinds of work we must not put them to. It stands in the way of forming clear intuitions about what is alienating labor for them—and thus likewise underlies the intuition that there seemingly is no such thing as alienating labor for them.

A robot's opacity is directly linked to its thin autonomy. Imagine standing next to a thinly autonomous robot in a room. The robot is turned off. You know it has the ability to scan the environment using the camera mounted as its "head", and to autonomously fulfill certain tasks given to it; e.g. to repair a machine for which we do not know the reason of its malfunction. Then you switch it on. The robot immediately starts to scan its surroundings by "looking around", i.e. by turning its camera-head left and right, up and down, also giving you the once-over. How does one experience this situation?

I think the natural reaction is a feeling of uncertainty—maybe alternating between positive suspense and a certain unease. Given the robot's thin autonomy, you are rather unsure of what exactly the robot is looking for, how exactly it is perceiving you (e.g. relying on which concepts and categories), and what the robot is "thinking". Will it push you aside as it identifies you as the reason for the machine's malfunction and wants to get you out of the way? Its opacity leads to a certain unpredictability.

The experience is largely analogous to that of facing a half-year-old baby who is looking about. Like a thinly autonomous robot, a half-year-old baby is not fully autonomous but constantly occupied with wringing structures and rules from its environment in a self-learning process. You know that the baby is scanning her environment and thinking something. But you do not exactly know what the baby is looking for or how she is processing information. Sometimes, the baby seems to be focusing on the "wrong" things, e.g. when she very intensely looks at a white wall for some prolonged period of time. Or when she is looking just by you rather than into your eyes, although you are very close to her. At the very least, you are unsure whether the baby is indeed taking in the "right" bits of information in the "right" way.

This experience does not drastically change if you are the designer of the robot (or the baby's parent), or if we made the robot's

calculations "transparent" by showing the operations it currently runs on some display. As the designer, you might have a clearer picture of what the robot is trying to do and in which categories it perceives its environment. But given the self-learning process it is running, you cannot exactly know what the robot is "thinking" at that very moment. This is also why a display giving us access to the inner life of the robot would not help us. First, as human beings we are not properly equipped to read code as it is executed. Second, even if we retraced every single line of code *AlphaGo Zero* used in beating a human opponent, this still would not allow us to understand its style of play.

No matter how well you know the self-learning robot (or the baby), you cannot fully grasp its "mind". It is in this sense that from thin autonomy follows a kind of opacity. I want to stress that I do not want to argue that there is a direct normative consequence of thinly autonomous robots' opacity for us.[12] I only mean to suggest that in light of their opacity, we cannot form clear intuitions about what would alienate them. Because we have no good epistemic access to their "thinking" and instrumental goal-setting, we are likewise blocked from fully understanding what we would need to do to further (or hinder) their self-realization and emancipation. When dealing with opaque beings, we are often lost with respect to the question of what we can do to help them—and thus likewise, with respect to the question of what would constitute active anti-emancipation on our part. This is why an agnostic stance as regards the question of whether we put robots to alienating labor is called for.

Because robots are not grown-up human beings, the epistemic access via introspection is blocked. Similarly, since robots are neither human beings, nor animals, but artifacts, we cannot rely on observing their natural behavior as a guide to what might be alienating for

[12] Levinas argues that there are drastic direct normative consequences of *alterity*—likewise an instance of us being unable to fully grasp another being's mind. Levinas holds that whenever we are facing an Other, the Other has all the rights, and we have all the duties (1969, 1985). I will not follow this route. First, because it has proven difficult to follow Levinas in claiming a higher status for the Other (e.g. Janicaud 1999). Second, we would need to argue that thinly autonomous robots are full-blown Levinasian Others (or at least very close to be)—of which I am not sure it can be done. Neither will I follow the route suggested by Derrida (1980, 2001: 90–1); namely, that with alterity comes dignity and a certain responsibility for the Other.

them. Thus, we face the question of whether reassigning Bostrom's paperclip maximizer or *Google*'s *AlphaGo* to some other task would be to alienate them—but without a reliable epistemic basis. To be sure, nothing feels wrong about reassigning them to do some entirely different job. But all we have precisely is this intuition, which is not particularly trustworthy.

One might try to argue that, since it is us who created them, and since we are those who decide what a certain thinly autonomous robot is designed to do, i.e. its function, we also get to decide what does or does not alienate it. This strikes me—at the very least—as an immodest thought. Bryson (2010) argues that less can go wrong with us arbitrarily using our power over robots than with us arbitrarily using our power over biological species.

> Perhaps unfortunately, we actually have almost as much control over other species and sometimes peoples as we do over robots. We as a culture do regularly decide how much and many resources (including space and time) we willingly allocate to others. But biological species hold exquisitely complicated and unique minds and cultures. If these minds and cultures are eliminated, they would be impossible to fully replicate. (Bryson 2010: 64–5)

In the case of robots, she goes on to say, we could basically just start anew if something goes wrong, without any real damage being done. But I think we should acknowledge that every time we develop a thinly autonomous entity with an ability for self-learning, we open up the door for a kind of evolutionary process which has emancipatory elements. Thinly autonomous entities learn and change. We built *AlphaGo Zero* precisely to do things better than we can. Thinly autonomous entities transcend our design efforts. This hardly strikes me as the position we want to be in when making claims about what work we may put them to without alienating them. A certain skepticism and the intent to err on the safe side seem more than called for.[13]

One might also try to argue that if a robot is a *universal* machine—i.e. that machine which has no one natural purpose, but which is a

[13] A stronger position with respect to this issue would be to hold that it already is normatively problematic for us to assign any function whatsoever to thinly autonomous entities. We could reject such "intrusions" of ours, for instance, against the background of fighting speciesism.

genuine universal-purpose tool—it is impossible to alienate a thinly autonomous robot to begin with. I think that argument might go through in the abstract. It certainly is not true for those robots we actually build. Concrete robots have a certain purpose which they are meant to fulfill, at least initially. This is not only due to the fact that we are, as of yet, incapable of building a fully autonomous AI. This is also due to practical reasons. We employ the robots we employ to fulfill particular purposes—so they are built to fulfill particular purposes.

To conclude, the intuition that we can put thinly autonomous robots to perform just any work (apart from immoral work) is influenced by the fact that, in light of their opacity, we have difficulties to tell what work we must not put them to. This leaves the question of what kind of work we must not put them to consciously open. (My intuitions on this matter are no better than everybody else's.) But it is important to stress that this lack of intuition on our part is risky for the robots. There is the constant threat that we alienate them, without us realizing that we do. Still, even if we do not have an intuitive answer to the question of what work exactly we must not put them to, we have the role-responsibility as employers not to alienate them. If we want to be sure not to be blameworthy with respect to how we discharge our role-responsibility as employers, we must not put them to perform just any work.[14]

Carl David Mildenberger
University of St. Gallen
Department of Philosophy
Unterer Graben 21
CH-9000 St. Gallen
Switzerland
carldavid.mildenberger@unisg.ch

[14] I would like to thank Dieter Thomä, Emmanuel Alloa, Michael Festl, Federica Gregoratto, Thomas Telios, Chris Paret, and Maria Dätwyler—as well as an anonymous referee—for many helpful comments. Earlier versions of this paper have been presented to audiences at the 2018 MANCEPT Workshops for Political Theory and in the colloquium for practical philosophy at the University of Zurich—and I greatly profited from the intense discussions there. I gratefully acknowledge support from the Research Commission of the University of St. Gallen (Grant No. 1031523), which helped me in preparing this work.

## *References*

Anderson, Elizabeth. 2017. *Private Government*. Princeton and Oxford: Princeton University Press.

Asaro, Peter M. 2011. A body to kick, but still no soul to damn: legal perspectives on robotics. In *Robot Ethics: The Ethical and Social Implications of Robotics*, ed. by P. Lin, K. Abney, and G. Bekey. Cambridge, MA: MIT Press.

Bostrom, Nick. 2014. *Superintelligence*. Oxford: Oxford University Press.

Bringsjord, Selmer; and Govindarajulu, Naveen Sundar. 2018. Artificial intelligence. In *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), ed. by Edward N. Zalta. https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/.

Bryson, Joanna J. 2010. Robots should be slaves. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. by Yorick Wilks. London: John Benjamins.

Capek, Karel. 1924. *R.U.R.: Rossum's Universal Robots*. Plamja: Praga.

Christman, John. 2018. Autonomy in moral and political philosophy. In *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), ed. by Edward N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/.

Deepmind. 2017. AlphaGo Zero: Learning from Scratch. https://deepmind.com/blog/alphago-zero-learning-scratch/.

Dennett, Daniel C. 1997. When HAL kills, who's to blame? Computer ethics. In *HAL's Legacy: 2001's Computer as Dream and Reality*, ed. by D. G. Stork. Cambridge, MA: MIT Press.

Derrida, Jacques. 1980. Violence and metaphysics. In *Writing and Difference*. Translated by Allan Bass. Chicago: University of Chicago Press.

Derrida, Jacques; and Roudinesco, Élisabeth. 2001. *De Quoi Demain...* Paris: Galilée.

Fox, Michael Allen. 2007. The moral community. In *Ethics in Practice*, ed. by Hugh LaFollette. Oxford: Blackwell.

Google. 2016. AlphaGo: Mastering the Ancient Game of Go with Machine Learning. https://ai.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html.

Gunkel, David J. 2014. A vindication of the rights of machines. *Philosophy and Technology* 27(1): 113–32.

Gunkel, David J. 2018. The other question: can and should robots have rights? *Ethics and Information Technology* 20: 87–99.

Hart, H. L. A. 1968. *Punishment and Responsibility*. New York: Oxford University Press.

Haydon, Graham. 1978. On being responsible. *The Philosophical Quarterly* 28(110): 46–57.

Hayek, Friedrich A. 1960. *The Constitution of Liberty*. Chicago: University of Chicago Press.

Jaeggi, Rahel. 2014. *Alienation*. Edited by Frederick Neuhouser. Translated by

Frederick Neuhouser and Alan E. Smith. New York: Columbia University Press.
Janicaud, Dominique. 1999. La Métaphysique d'Emmanuel Levinas. *Noesis* 3: 12–35.
Kamm, Frances M. 2007. *Intricate Ethics*. Oxford: Oxford University Press.
Leopold, David. 2018. Alienation. In *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), ed, by Edward N. Zalta. https://plato.stanford.edu/archives/fall2018/entries/alienation/.
Levinas, Emmanuel. 1969. *Totality and Infinity: An Essay on Exteriority*. Translated by Alphonso Lingis. Pittsburgh: Duquesne University Press.
Levinas, Emmanuel. 1985. *Ethics and Infinity: Conversations with Philippe Nemo*. Translated by Richard A. Cohen. Pittsburgh: Duquesne University Press.
Martin, Mike W.; and Schinzinger, Roland. 2004. *Ethics in Engineering*. New York: McGraw-Hill.
Marx, Karl. [1867] 1992. *Capital: Volume 1: A Critique of Political Economy*. Edited by Ben Fowkes. London: Penguin Books.
Marx, Karl, and Friedrich Engels. [1845] 1975. The Holy Family. In *Karl Marx, Friedrich Engels: Collected Works 4*. London: Lawrence and Wishart.
Omohundro, Steve. 2008. The basic AI drives. *Proceedings of the First AGI Conference*. http://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/.
OpenAI. 2017. More on Dota 2. https://blog.openai.com/more-on-dota-2/.
Pearl, Judea. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. https://arxiv.org/abs/1801.04016.
Poel, Ibo van de; and Royakkers, Lambèr. 2011. *Ethics, Technology, and Engineering*. Malden: Wiley-Blackwell.
Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417–57.
Silver, David; Schrittwieser, Julian; Simonyan, Karen; Antonoglou, Ioannis; Huang, Aja; Guez, Arthur; Hubert, Thomas; *et al.* 2017. Mastering the game of Go without human knowledge. *Nature* 550: 354–9.
Sparrow, Robert. 2007. Killer robots. *Journal of Applied Philosophy* 24(1): 62–77.
Stevens, Robert. 1988. Coercive offers. *Australasian Journal of Philosophy* 67: 472–5.
Waller, Bruce N. 1993. Responsibility and the self-made self. *Analysis* 53 (1): 45–51.
Waller, Bruce B. 2011. *Against Moral Responsibility*. Cambridge, MA: The MIT Press.
Williams, Bernard; and Smart, J.J.C. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.
Zimmerman, David. 1981. Coercive wage offers. *Philosophy and Public Affairs* 10: 121–45.