Philosophical Studies
https://doi.org/10.1007/s11098-023-02027-0

# Getting machines to do your dirty work

Tomi Francis[1] · Todd Karhu[2]

Accepted: 28 July 2023

## Abstract

Autonomous systems are machines that can alter their behavior without direct human oversight or control. How ought we to program them to behave? A plausible starting point is given by the Reduction to Acts Thesis, according to which we ought to program autonomous systems to do whatever a human agent ought to do in the same circumstances. Although the Reduction to Acts Thesis is initially appealing, we argue that it is false: it is sometimes permissible to program a machine to do something that it would be wrong for a human to do. We advance two main arguments for this claim. First, the way an autonomous system will behave can be known in advance. This knowledge can indirectly affect the behavior of other agents, while the same may not be true at the time the system actually executes its programming. Second, a lack of knowledge of the identities of the victims and beneficiaries can provide a justification during the programming phase that would be unavailable to an agent at the time the autonomous system executes its programming.

**Keywords** Autonomous systems · AI ethics · Identified vs statistical lives · Decomposition test · Rational irrationality

Autonomous systems—machines such as self-driving cars and autonomous weapons systems which can alter their behaviour without direct human oversight or control—can be expected to become increasingly prevalent over the next several decades. Accordingly, it is increasingly urgent to find out which moral principles

✉ Todd Karhu
todd.karhu@kcl.ac.uk

Tomi Francis
tomi.francis@philosophy.ox.ac.uk

[1] Global Priorities Institute, University of Oxford, Trajan House, Mill Street, Oxford OX2 0DJ, UK

[2] The Dickson Poon School of Law, King's College London, Somerset House East Wing, London WC2R 2LS, UK

Published online: 09 September 2023

should guide the way that they are programmed.[1] A natural suggestion is that we ought to program them to do whatever it is that *we* ought to do in the same circumstances. That is, we might accept the

> *Reduction to Acts Thesis* It is permissible to program an autonomous system to perform act type A in choice situation C if and only if an impartial human agent would be permitted to perform act type A in choice situation C.[2]

If the Reduction to Acts Thesis is true, then any question of how an autonomous system ought to be programmed to behave in a given choice situation can be reduced to a question about what a human agent ought to do in the same choice situation. To work out how we ought to program a self-driving vehicle to act in a collision situation, we can simply ask how an impartial human agent should direct it to act in that situation. However, we shall argue that the Reduction to Acts Thesis is false. Specifically, we shall argue that it can be morally permissible to program an autonomous system to act in a way that would be impermissible for a human. Sometimes, we should get machines to do our dirty work.[3]

## 1 Moral machines?

The Reduction to Acts Thesis seems to be supported by the following line of reasoning. We ought to program autonomous systems to do what they ought to do. What they ought to do is no different to what a human being facing the same choice situation ought to do. After all, no matter whether it is a human or an autonomous system facing a choice, what ought to be done is determined by weighing up the morally relevant effects of the available alternatives. The morally relevant effects of the alternatives will be the same in both cases, since this is just what it means for the human and the autonomous system to face the same choice situation. Therefore, we ought to program autonomous systems to do whatever it is that we ought to do.

This way of arguing for the Reduction to Acts Thesis is seductive, but ultimately unsuccessful. Its most conspicuous flaw is that it tacitly assumes that we can

[1] We shall skim over one potential complication: in some cases, an autonomous system may be programmed by one party, but deployed by another party in a way not intended by the original programmer. For example, an autonomous weapons system might be programmed to select targets liberally, on the explicit understanding that the system is only to be used in locations which are clear of noncombatants; if the system is subsequently deployed by the military in an area containing many noncombatants, and the weapons system subsequently kills some of these noncombatants, the fault presumably lies with the military, rather than the programmers. (This example is adapted from Sparrow 2007: 69). We shall restrict our attention to cases in which the autonomous system is acting in a way that is intended by the programmer.

[2] We here take *act types* to be individuated by the morally relevant effects they produce, and we understand *choice situations* to be sets of act types. For instance, if there are two possible situations in which a human and an autonomous system can, depending on their behaviour, either save the same person from the same harm or not, then the human and the autonomous system face the same choice situations, even though of course the act tokens available to them are distinct.

[3] This phrase is due to Johann Frick (2015: 211).

sensibly talk about what autonomous systems "ought" to do. But this kind of talk is misleading at best. Autonomous systems of the sort we can produce today are *not* the sorts of entities which have moral obligations in the same sense as we do. They cannot be made better or worse off, they do not have special obligations or agent-centred prerogatives, they are not morally responsible for their behaviour, and they are not capable of having or responding to moral duties. In short, autonomous systems are not moral agents.[4]

There may be some sense in which autonomous systems ought to do things. Judith Jarvis Thomson (2008) claims that it makes perfect sense to say that machines ought to do things even when they clearly lack anything approaching autonomy or cognition: a toaster, she suggests, ought to toast bread.[5] On her view, a toaster which didn't toast bread would be a defective member of its kind. But if this is what we mean when we say that autonomous systems ought to do things, the claim that what autonomous systems ought to do is the same as what *we* ought to do becomes very implausible. An autonomous weapon system ought to fire, in Thomson's sense, if it receives a legitimately transmitted instruction to fire from a member of the military who is authorised to operate the autonomous system. This is the case even if the target of the autonomous weapons system is an innocent civilian. But an impartial human operator plainly ought *not* to direct the weapons system to fire at an innocent civilian, even if she receives a legitimately transmitted instruction to fire. So it is false that what an impartial human operator ought to do is no different to what an autonomous system ought to do, when "ought" is understood in Thomson's sense.[6]

## 2 The extrinsic effects of programming

Suppose you program an autonomous vehicle to brake as sharply as possible, if it detects that it would otherwise collide with a pedestrian. What you have done is not, in general, equivalent to causing it to brake in one specific instance. The vehicle in question might never find itself in that collision situation, and so never execute its programmed response. Or, it might find itself in multiple collision situations of that type, and so execute its programmed response multiple times. In neither case do you cause the autonomous vehicle to brake once.

What you really do is give the autonomous vehicle the *disposition* to brake sharply when it detects that it would otherwise collide with a pedestrian. More generally, to program an autonomous system to behave in a certain way under certain conditions is not to produce the relevant action under those conditions. It is rather

[4] For further arguments for this claim, see Purves et al. (2015: 861) and Talbot et al. (2017: 258–260).

[5] This position appears to enjoy support from outside the philosophical community. According to familycuisine.net (2021), which we take to be an authoritative source on the matter, "a great toaster ought to toast bread evenly and persistently".

[6] Even if there are other senses of "ought" to which autonomous systems might be susceptible, they won't have the same sense as the "ought" of moral obligation as it applies to humans. Thus, there is no reason to expect that what an autonomous system "ought" to do is no different to what a human being has a moral obligation to do.

to dispose the autonomous system to behave in that way under those conditions. The root of our objections to the Reduction to Acts Thesis is that the morality of programming is much closer to the morality of producing and acquiring dispositions than it is to the morality of acting directly. The Reduction to Acts Thesis fails because there are cases where the two do not line up.

One way in which they fail to line up has been discussed extensively by Derek Parfit (1984). Parfit was concerned with the rationality of acquiring dispositions in cases in which we are *transparent*: when our dispositions are visible to others. In such cases, he argued, we can be better off if we are disposed to do what would be worse for ourselves; thus, rationality can require us to become disposed to act irrationally. Analogously, when autonomous systems are transparent in the same way, it can be permissible to program them to act in a way that would be impermissible.[7] An example might help bring this out.

Suppose that in a certain country in fifty years' time a police department is planning to replace its hostage negotiators with robots. One option for them is to program these robots to immediately respond with lethal force to any hostage situation, predictably resulting in the deaths of the hostages as well as their kidnappers. The way the robots are programmed could be made public knowledge, so that would-be kidnappers would know that they will almost certainly die if they attempt to take hostages. By programming the robots in this way, one might in all likelihood prevent any kidnappings from taking place. Thus, it may well be permissible to programme the robots to use excessive force. However, it seems impermissible for a human operator to manually direct a robot to respond with such indiscriminate force in an actual hostage situation. If that is right, we have a counterexample to the Reduction to Acts Thesis.

The same phenomenon gives rise to other counterexamples. It is wrong to destroy the world, even if nuclear weapons have just been launched at your country. But if you can publicly program an autonomous system to destroy the world if nuclear weapons are launched, then provided rival states are certain to behave rationally and there is no chance of a mishap, the effect of doing so would be to eliminate the possibility of nuclear war. There may be some moral objection to programming the the doomsday device, even when you can be certain nothing will go wrong. But given the effects, programming it is almost certainly permissible, all things considered.[8]

More generally, the decision to program an a machine in a certain way might itself have morally significant *extrinsic effects*, over and above the effects caused by the operation of the autonomous system. These extrinsic effects can render it permissible to dispose a machine to do things that it would be impermissible for a human agent to do.

[7] We also think that it can be impermissible to program them to act in a way that would be permissible, for similar reasons.

[8] On the point that, at least in cases of certainty, the threat of nuclear retaliation is surely less serious than actual retaliation, see Williams (1982).

## 3 The argument from ex ante interests

Our arguments so far leave open that the Reduction to Acts Thesis is correct when restricted to cases in which programming has no morally significant extrinsic effects. In fact, however, we doubt that even this restricted principle is correct. Producing a disposition can sometimes have morally significant extrinsic effects. But there is another important difference between the morality of directly acting and the morality of producing dispositions. This is that one is often in a different epistemic position at the time of producing the disposition than one would be when acting on it. In the case of autonomous systems, programmers are often in a worse epistemic position than an impartial human would be in the relevant choice situation.

One reason this might be thought to challenge the Reduction to Acts Thesis has to do with uncertainty about whether a given choice situation will arise. Talbot et al. (2017) provide an argument along these lines. They suggest that the strength of the moral constraint against imposing possible harm might vary non-linearly with the probability that the harm will occur, but that the strength of our reason to provide benefits does vary linearly with their probability. On their suggestion, it might well be impermissible to kill one person to relieve any number of other people's headaches, but permissible to inflict a 0.01% chance of death on a person to obtain a 0.01% chance of curing a billion headaches. Thus, it might be permissible to program an autonomous system to kill one person to relieve a billion people's headaches, provided there were only a 0.01% chance that this choice situation would actually arise.

If the constraint on imposing possible harm really does vary non-linearly with the probability of its imposition while reasons to benefit do not, then this argument establishes that the Reduction to Acts Thesis is false. However, it would be surprising if our reasons to avoid harming were in this way structurally different from our reasons to benefit. At the very least this is a fairly controversial premise even among those who maintain that there is no number of headaches to be cured which can justify taking a life.[9] It is also worth noting that the argument, if successful, applies to a more limited class of cases than it might at first seem. Talbot et al. suggest that it might often be justified to ignore constraints when programming autonomous vehicles, as they "will very rarely get into situations where respecting rights and maximizing utility conflict" (2017: 265). That seems right if you're programming a single vehicle in isolation: if there is a 0.01% chance of a given vehicle getting into a certain type of collision situation, then you might be able to ignore constraints when deciding how to program the vehicle to behave in that situation. But if you're programming for a company which expects to sell 100, 000 vehicles, you'd better not

[9] Lazar and Lee-Stronach (2019) provide a decision theory for "absolutist" moral theories, on which there are some moral considerations (like killing) which are weightier than any number of lesser moral considerations (like preventing headaches). Their decision theory obeys the axioms of Expected Utility Theory. In particular, it satisfies Independence, which rules out that the strengths of moral constraints can vary non-linearly with their probabilities in the way that Talbot et al. require.

ignore constraints: the probability that no such collision situation will arise will be roughly $0.9999^{100,000}$, or about 0.0045%.

There is another way in which programmers' epistemic limitations can challenge the Reduction to Acts Thesis. This challenge has to do not with uncertainty about whether a given choice situation will arise, but with uncertainty about who will benefit and who will be harmed when an autonomous system performs a certain act type. The idea is that optimific harms will often be justifiable to each person ex ante when each person has an equal chance of being a victim or beneficiary before the choice situation arises. In such cases, there will be an important justification for programming the autonomous system to carry out the optimific harm, even if it would, on common-sense morality, be impermissible to carry out the optimific harm at the time. We can bring this out with the following scenario:

> *Medical Robot* A rare disease is afflicting three unconscious patients: Ali, Bianca and Annabelle, who have blood types A, B and AB respectively. They are hooked up to a medical robot, which is designed to diagnose the progress of the disease and treat it. At some point, the disease will progress, developing into one of three stages, with equal probabilities: the A stage, the B stage, or the AB stage. All three patients' diseases will progress to the same stage, whatever that may be. Each stage will be harmless for anyone with a blood type of the same name, but it will be deadly for anyone else if untreated. After the disease has progressed, it will be possible to administer a drug either to all three patients or to none of them. If administered, this drug will cause the disease to behave differently: it will now kill whoever has the blood type which would previously have ensured survival, while the others will make a full recovery.

Consider first *Case One*, in which a human being is directing the medical robot after the disease has progressed, at which point it is known which person will die after taking the drug. In that case, administering the drug amounts to knowingly killing one healthy patient to save two who would otherwise die. Most people will balk at this, even many of those who believe that it is permissible to turn a lethal train away from five and toward one.[10]

Next consider *Case Two*. As before, a human being is directing the medical robot, and the disease has already progressed. This time, however, it is unknown whether the disease has progressed to the A, B or AB stage. If a decision must be made right away, it seems permissible to direct the machine to administer the drug to all three patients, since this will reduce each patient's chance of death from 2/3 to 1/3.

Finally, consider *Case Three*, in which the disease has not yet progressed, and an irreversible decision must be made now about whether to program the medical robot to administer the drug upon progression of the disease. At the point at which the drug is ready to be administered, it will be known whether the disease is in the

[10] Philippa Foot and Judith Jarvis Thomson, for example, both deem it impermissible to activate a machine which will release a life-saving gas into a room containing five dying patients, knowing that the lethal fumes from the machine will necessarily travel into the different room of a single healthy individual, killing her. (See Foot 1967: 14 and Thomson 1985: 1407).

A, the B, or the AB stage (for all three patients). But we do not know this *now*. We are therefore in the same position as in Case Two: programming the medical robot to release the drug reduces each patient's chance of death from 2/3 to 1/3. Again, it seems permissible to do so.

If it is permissible to administer the drug in Case Three, but impermissible to administer the drug in Case One, we have a counterexample to the Reduction to Acts Thesis. The claim that it is permissible to administer the drug in Cases Two and Three might be doubted. It is true that, because we do not know who will be killed by the drug, administering the drug is in everyone's ex ante interest in each case. Still, we know we will be killing one innocent person to save two. Is that not enough to say that administering the drug is wrong?

Proponents of the Reduction to Acts thesis can do better than ask rhetorical questions. They can appeal to principles like the

> *Principle of Full Information* When one knows that, in every state of the world with positive probability, one would rightly rank two alternatives in a particular way, then one should so rank them.
>
> Fleurbaey and Voorhoeve (2013: 120)

According to the Principle of Full Information, if it is impermissible to administer the drug in Case One, it must also be impermissible to administer the drug in Case Two (and presumably Case Three as well). And the Principle of Full Information is plausibly a requirement of rationality. So, doesn't this mean that administering the drug in Case Two is indeed wrong? Not so fast. Considered in isolation, the Principle of Full Information is compelling as a requirement of rationality. But it is much less obvious that we should assume it *given* the correctness of the sort of common-sense moral judgements we are assuming.[11]

Moral common sense suggests that we ought to prevent a single death instead of preventing millions of headaches.[12] But the combination of this claim with the Principle of Full Information is not plausible. Millions of mild painkillers like aspirin are consumed each day in the developed world to relieve minor pains like headaches. The increased traffic from the delivery of these drugs for non-essential purposes is bound to cause at least a few deaths. So, if we think we ought to prevent one death over any number of headaches, we can be virtually certain that we would rank the alternative in which the additional drugs are not delivered over the alternative in which they are delivered. Yet few people would seriously believe that we should stop delivering aspirin for non-essential purposes, even if we could be confident that the only morally relevant factors at play were the deaths and the headaches.[13]

[11] Utilitarians, of course, won't accept the claim that it would be wrong to kill the healthy patient in *Case One* in the first place. An important argument due to John C. Harsanyi shows that, given a few other minor assumptions, it is *only* Utilitarians who can consistently accept both principles like the Principle of Full Information and the claim that we should do whatever is in everyone's ex ante interest. See Harsanyi (1955), Broome (1991) or Fleurbaey (2009) for details.

[12] In fact, Voorhoeve is one of the main proponents of this view; see Voorhoeve (2014, 2017).

[13] Norcross (1997: 159–167) uses a similar example to argue against the claim that we ought to prioritise preventing deaths over preventing any number of headaches. Contra Norcross (161), it seems to us plau-

Moreover, a direct argument can be given for the claim that we should administer the drug in Cases Two and Three.[14] Consider *Case Four*, in which a human being is directing the medical robot to administer the drug, as in *Case Two*. This time, however, there are three drugs rather than one. Each drug does exactly the same thing as the drug administered in Cases One through Three, with the exception that the first drug affects only Ali, the second affects only Bianca, and the third affects only Annabelle. As before, the disease has already progressed, but it is unknown whether it progressed to the A, B or AB stage. If all three drugs are administered, the situation will therefore be exactly the same as if the drug is administered in Case Two.

We can now ask: Should a human operator direct the medical robot to administer Ali's drug? Yes. The drug will significantly decrease Ali's probability of death, from 2/3 to 1/3, without affecting anybody else. There is, of course, one potentially morally relevant factor to beware of: if Ali dies due to taking the drug, then the operator will have killed him, whereas if Ali dies due to not taking the drug, the operator would have merely let Ali die. Even so, it does not seem that this factor should outweigh the drastic improvement in Ali's survival prospects. To see this, imagine a different scenario: Ali has cancer, which will kill him with 2/3 probability, but the cancer can be cured by a dangerous surgery, which runs the risk of killing him. If Ali is unconscious, would it be permissible to operate on him? Maybe not, if his of risk of death from surgery is only slightly smaller than his risk of death from cancer. But if the risk of death from surgery is 1/3, it would clearly be permissible to perform it. It would be absurd to allow Ali to face an extra one third risk of death just to avoid the possibility that he dies due to a botched operation rather than cancer.

If it is permissible to operate on Ali in the cancer case, then it is permissible to administer his drug in Case Four. And what goes for Ali goes for the others: nothing in the preceding argument depended on the fact that one had *not* already given the tailored drugs to the other patients. (If one operates on sufficiently many cancer patients, sooner or later at least one of them will die. But one should continue operating so long as doing so dramatically improves the survival chances of each individual.) So, regardless of whether one has given tailored drugs to the others, it is permissible to give the tailored drug to Ali, and it is permissible to give the tailored drug to Bianca, and it is permissible to give the tailored drug to Annabelle. Surely, then, it is permissible to give all of the tailored drugs at once. But that amounts to the same thing as administering the drug in Case Two. So it must be permissible to administer the drug in Case Two. And, if it is, it must be permissible to program the medical robot to later administer the drug, as in Case Three.

As we have seen, the conclusion of this argument is inconsistent with the Principle of Full Information, provided that, as common sense suggests, it would be impermissible to administer the drug in Case One. Note, also, that the argument does not beg the question against the Principle of Full Information. When we

Footnote 13 (continued)
sible (though arguable) that identified deaths outweigh any number of headaches, while statistical deaths do not.

[14] This argument is based on Hare (2016).

consider administering each tailored drug, we do *not* know that we would rank the alternative, namely not administering this tailored drug, as the better alternative in every state of the world. Given that the tailored drug would save the patient's life, it is better for the drug to be administered; given that administering the tailored drug would result in the patient's death, it is better for the drug not to be administered. Since we do not know which state of the world actually obtains, the Principle of Full Information is silent in cases like these, as Fleurbaey and Voorhoeve (2013: 120–121) acknowledge.

We conclude that it is permissible to administer the drug in Case Three. Moral common sense suggests that it is wrong to administer the drug in Case One. These two judgements together constitute a counterexample to the Reduction to Acts Thesis. So philosophers in the business of respecting common-sense morality had better reject the Reduction to Acts Thesis.

One might wonder how far arguments from ex ante interests go. On one extreme, one might believe that *whenever* an act is in everyone's ex ante interest, it is permissible to perform this act. Some authors, such as Gogoll and Müller (2017), seem to be happy to accept this conclusion. Others disagree.[15] We take no stance on this question. Our argument requires only that an action's being in everyone's ex ante interests affords it *some* justification.

# 4 Objections and replies

## 4.1 The Decomposition Test

The counterexamples we have given to the Reduction to Acts Thesis are supposed to show that sometimes it is permissible to get a machine to do something that would ordinarily be wrong. This might seem suspicious. If an action will be unjustifiable to others at a later time, how can we evade the demand for justification by getting a machine to do our dirty work for us? Johann Frick (2015) has made this worry more precise. He claims that we should accept the

> *Decomposition Test* If a rule or procedure can be decomposed into a sequence of distinct causal stages, each of which involves the voluntary action of some agent (or of a surrogate for human agency, such as a programmed machine), then it is permissible to adopt and act on this rule or procedure only if the actions it requires at every stage are justifiable to each person *at that time*.
>
> Frick (2015: 205)

[15] To take two examples: Mogensen and MacAskill (2021:10) point out that it would intuitively be wrong for a community could install an AI-controlled braking system in their ambulances that would make it impossible to prevent the ambulances from running over innocent pedestrians when not stopping is necessary to save a greater number of people in critical condition on board the ambulance. (Theirs is a variation on Kamm's 1996: 303 Ambulance Case.) And Hübner and White (2018) note that under certain conditions it would be in everyone's ex ante interest to program an autonomous system to select and kill healthy people when it could to use their organs to save several lives, an idea most people would find abhorrent—this case originally appears in Harris (1975).

Recall Case Three of our Medical Robot scenario, in which we must now decide whether to irreversibly program the medical robot to administer the drug to all three patients, after it becomes known which of them stands to die. The Decomposition Test suggests that doing so would be impermissible. To program the medical robot to later administer the drug is to adopt a procedure in which a surrogate for human agency will kill one to save two, which is presumably unjustifiable at the time.

But is it really appropriate to apply the Decomposition Test to an act performed by programmed machines? We think not, broadly for the reasons given in Sect. 1. Machines are not moral agents, so to say that their behaviour is justified or unjustified looks to us like a category mistake. However, Frick believes that the Decomposition Test must apply to programmed machines if we are to get the right results in cases like

> *Automatic Experiment* One hundred children suffer from paraplegia. A doctor can either:

(A) do nothing, or

(B) initiate an automatic process which selects ten children via a fair lottery, conducts lethal medical experiments on them, and then uses the knowledge gained to cure the remaining 90 children.

Frick thinks that it would be impermissible to do (B) in *Automatic Experiment*, even though this would be in every child's ex ante interest.[16] At the same time, he believes that it would be permissible to give each of the hundred children a risky pill which would predictably kill ten of them and cure the rest of their paraplegia, provided the pill would give to *each child* a 90% chance of survival. According to Frick, the difference between the two cases comes down to the fact that *Automatic Experiment* involves allowing a machine—a surrogate for human action—to do something that would be unjustifiable to ten children *at the time*. To support this, he points out that it would be impermissible for a human doctor to go through with the lethal medical experiments, even if the ten unlucky children were previously selected by a fair lottery. He says that the killing of these children is

> unjustifiable, not because it is carried out by a human agent rather than a machine, but because it avoidably places the uncompensated and severe burden of death on 10 children in order to cure 90 others from the less serious burden of paraplegia. And this feature, of course, carries over to Automatic Experiment.
> (Frick 2015: 211)

[16] We assume here that it is better in expectation for each child to have a 90% chance of survival without paraplegia, and death otherwise, than to have certainty of survival with paraplegia. If you doubt this, the example could be adjusted so that more children survive, or paraplegia could be switched with some even worse affliction.

It is true in a sense that whether a killing is carried out by a machine rather than by a human agent makes no moral difference: it is no more justifiable to kill a person by directing a drone to kill them than it is to kill that person with one's bare hands. But there is another sense in which a killing by a human can be unjustified, while an equivalent killing by a machine is not. For the question of justification arises at different times. When a machine kills somebody, we can ask whether it was justified to program the machine in the way that it was programmed; to deploy it in the way that it was deployed; and so on. But once all the associated human actions are accounted for, no further question of justification arises. In particular, there may be no human action *at the time of the killing* which requires justification. It is not that if a machine cannot be interfered with, its killing someone may be justified at that time. Rather, there is nothing susceptible of being justified or unjustified at that time. In the case of Automatic Experiment, the imposition of death on the ten children after they are selected is not the sort of thing that calls for justification. What calls for justification is only the choice of (B): the initiation of the automatic process that leads to the imposition of death on ten as-yet unknown children.

It may yet be impermissible to choose (B) in *Automatic Experiment*.[17] But, if so, this is not because the Decomposition Test can be applied to programmed machines. Rather, it is because there is something especially wrong about killing children in order to use their bodies for medical experiments. To see this, consider the following variation on Automatic Experiment:

> *Postmortem Experiment* One hundred children suffer from paraplegia. A doctor can either:

(A) do nothing, or
(B) press a button which randomly kills ten of the children. Once those children are dead, medical experiments can (and will) be conducted on their bodies, which will allow the remaining 90 children to be cured. None of the children object to their bodies being used for medical experiments after death.

*Postmortem Experiment* looks a lot like *Automatic Experiment*. There is one main difference: at no point does it involve levelling a serious burden on fewer "identified" individuals in order to prevent a less serious burden from befalling a greater number of "statistical" individuals. When the button is pressed, ten children immediately die; after that, given that none of the children object to their bodies being experimented upon, it seems clearly justifiable *at that time* to conduct the medical experiments in order to cure the remaining 90 children (even if it is wrong to press the button in the first place).

If the wrongness of performing (B) in *Automatic Experiment* were explained by its failing the Decomposition Test, then (B) would be wrong *because* the mechanism

[17] A case with a similar structure to *Automatic Experiment* is discussed by Lenman (2017: 102–104).

would carry out the killings at a time at which they would not be justifiable to each person. Since option (B) in *Postmortem Experiment* lacks this feature, we would expect it to be permissible: pressing the button can be justified to each person as being in their ex ante interest, and nobody objects to their body being used after their death to help the remaining 90 children. But this is not what we find. It seems clear to us that *if* doing (B) is impermissible in *Automatic Experiment*, then doing (B) is impermissible in *Postmortem Experiment*.

It therefore seems to us that the Decomposition Test plays little to no role in the impermissibility of choosing (B) in *Automatic Experiment*. The Decomposition Test is not needed to explain why doing (B) is wrong in either Postmortem Experiment or Automatic Experiment, and so the wrongness of doing (B) in Automatic Experiment cannot be used as an argument for the Decomposition Test. We thus see little reason to think that it is appropriate to apply the Decomposition Test to the case of programmed machines.

### 4.2 The circumvention problem

Recall the *Medical Robot* case from Sect. 3. We argued there that it would be permissible to program a medical robot to administer a drug which will kill one of three people and save the lives of the other two, provided that at the time of programming, each person would have a 1/3 chance of being killed and a 2/3 chance of being saved. We also claimed that it would be impermissible to administer this drug if it were known who the victims and beneficiaries would be. This seems to lead to the apparently paradoxical result that it would be right to program the machine to administer the drug, but it would later be right to circumvent this original programming in order to prevent administration of the drug.

We cannot reply to this objection by claiming that the circumstances in which denying the Reduction to Acts Thesis would force us to be irrational are unlikely to arise in practice. That may well be true, but the problem remains unless we can rule out all such cases on principle.[18] Surely it is irrational to program a machine to do something, only to foreseeably later cause it to circumvent its programming. It cannot be irrational to follow the correct moral theory. Therefore, the correct moral theory cannot tell us to do such things.

Here's a better response. We might restrict our claims so that the Reduction to Acts Thesis is only false in cases in which future circumvention of programming is or can be rendered impossible. (Indeed, we made this restriction already in *Case Three* of our Medical Robot scenario.) There is a good rationale for making this restriction. When it comes to sequential decision-making, plausibly we should take into account what we know we will do in the future. Applying foresight in this way,

[18] Our reasons to program an autonomous system a certain way are also reasons to ensure that we will later not interfere with the operation of this autonomous system. We therefore have reasons to make circumvention impossible whenever such cases arise (though there may be practical considerations that tell against doing this). However, there will sometimes be cases when we are unable to prevent our later circumventing the operation of the autonomous system.

if we know that we will later be in a position to circumvent the programming of an autonomous system, we will never in the first program it to do something it would be wrong for a human to do, since we know in advance that it would be fruitless to do so.

A final potential response is to maintain that if we may program a machine to do what would be the "wrong" thing taken by itself, this very fact makes it permissible to avoid circumventing this programming later on.[19] This thought has more going for it than it might at first appear. Consider the distribution of a given stock of resources among ten people, none of whom is in a position to give or withhold consent to the distribution. Suppose that if we divide the resources equally among these ten people, each person will receive ten units. If we divide the resources unequally, eight of the people will have twelve units of wellbeing, while the remaining two people will each have five units of wellbeing. If we think it is important to give each person a greater expected benefit, we might decide to distribute the resources unequally, but give each person a fair chance by holding a lottery. But drawing the lots does not magically give each person their allotted share. After the drawing, we face the choice of dividing the resources according to the lottery or, alternatively, dividing them equally despite the lottery. Plausibly, if we *hadn't* had a lottery, it would be best to now divide the resources equally. But given that we *have* had a lottery, we should divide them unequally according to the result of the lottery.[20] In much the same way, if we program a machine in a way that is justifiable to each person ex ante, this might permit us not to interfere with its programming later, even though its behaviour would not be justifiable to each person ex post. And perhaps this might be true even if somebody else has programmed the machine, in the same way that it may be justifiable to unequally divide resources in line with a fair lottery held for this purpose by somebody else.

## 5 Concluding remarks

The Reduction to Acts Thesis, if true, would provide a simple schema for approaching questions about how autonomous systems ought to be programmed to behave in a given choice situation: first, ask how we ought to behave in the same choice situation, then program the autonomous system to behave in this way whenever this choice situation arises. However, since the Reduction to Acts Thesis is false, this simple schema is liable to mislead us. To avoid being misled, we must be careful not to reason as though the identities of the victims and beneficiaries of an autonomous system's behaviour are always known at the time of programming. And, when the way an autonomous system is programmed to behave might come to be known by agents, we must not forget that this knowledge may well affect what these agents will do.

[19] That is, we might instead employ "resolute choice"; see McClennen (1985, 2000). For criticism, see Gustafsson (2022: Ch. 7).

[20] A similar point has been made by Broome (1990–1991a).

What, then, is the correct account of the morality of programming autonomous systems? We claimed in Sect. 2 that the morality of programming machines is more akin to the morality of producing dispositions than it is to the morality of acting directly. This suggests an alternative to the Reduction to Acts Thesis, namely the

> *The Reduction to Dispositions Thesis* It is permissible to program an autonomous system to perform act type A in choice situation C if and only if an impartial human agent facing the same pattern of choice situations would be permitted to acquire a disposition to perform act A in choice situation C.

Should we accept the Reduction to Dispositions Thesis? Perhaps, but there are some reasons to doubt it. One might reasonably think that part of living a good life is acting in accordance with morality. If so, then acquiring an immoral disposition is in one way bad for us, whereas no similar consideration applies to programming a machine. One also might think it could be impartially bad for there to be immorally disposed people in the world. Perhaps the world would be worse a worse place if you were to acquire a disposition to torture llamas, even if you will never encounter one. Alternatively, it might make the world worse when *people* act wrongly, but not when *autonomous systems* "act wrongly".[21]

These reasons for doubt aren't decisive, because it's not clear whether these are genuine ways in which causing people to be disposed to behave immorally would be bad. It therefore seems to us that the jury is out on the Reduction to Dispositions Thesis. But, at any rate, the Reduction to Acts Thesis is false. Sometimes, the right thing to do is to get a machine to do the wrong thing for you.

**Acknowledgements** For valuable feedback, we would like to thank Henrik Kugelberg, Seth Lazar, Aidan Penn, Rob Reich, Charlotte Unruh, and participants at the the Oxford Institute for Ethics in AI research seminar, the Stanford HAI weekly seminar, and the Stanford Center for Ethics workshop.

## Declarations

**Conflict of interest** The authors declare none.

[21] On the claim that it is in itself bad that people act wrongly, see Parfit (2017: 354–357, 400–406).

## References

Broome, J. (1990–1991a). Fairness. *Proceedings of the Aristotelian Society*, *91*(1), 87–102.

Broome, J. (1991). *Weighing goods: Equality, uncertainty and time*. Oxford: Basil Blackwell.

familycuisine.net. (2021). The best toasters we tested in 2021. https://familycuisine.net/whats-the-best-toaster/. Accessed 2022-03-08.

Fleurbaey, M. (2009). Two variants of Harsanyi's aggregation theorem. *Economics Letters, 105*(3), 300–302.

Fleurbaey, M., & Voorhoeve, A. (2013). Decide as you would with full information! an argument against ex-ante pareto. In N. Eyal, S. A. Hurst, O. F. Norheim, & D. Wikler (Eds.), *Inequalities in Health: Concepts, Measures, and Ethics, chapter 8* (pp. 113–128). New York: Oxford University Press.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *The Oxford Review, 5*, 5–15.

Frick, J. (2015). Contractualism and social risk. *Philosophy & Public Affairs, 43*(3), 175–223.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics, 23*(3), 681–700.

Gustafsson, J. E. (2022). *Money-pump arguments*. Cambridge: Cambridge University Press.

Hare, C. (2016). Should we wish well to all? *The Philosophical Review, 125*(4), 451–472.

Harris, J. (1975). The survival lottery. *Philosophy, 50*(191), 81–87.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy, 63*(4), 309–321.

Hübner, D., & White, L. (2018). Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice, 21*(3), 685–698.

Kamm, F. M. (1996). *Morality, Mortality Volume 2: Rights, Duties and Status*. New York: Oxford University Press.

Lazar, S., & Lee-Stronach, C. (2019). Axiological absolutism and risk. *Noûs, 53*(1), 97–113.

Lenman, J. (2017). Contractualism and risk imposition. *Politics, Philosophy & Economics, 7*(1), 99–122.

McClennen, E. F. (1985). Prisoner's dilemma and resolute choice, chapter 5. In R. Campbell & L. Sowden (Eds.), *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's problem* (pp. 94–104). Vancouver: University of British Columbia Press.

McClennen, E. F. (2000). The rationality of rules, chapter 2. In J. Nida Rümelin & W. Spohn (Eds.), *Rationality, rules, and structure* (pp. 17–33). Berlin: Springer.

Mogensen, A., & MacAskill, W. (2021). The paralysis argument. *Philosophers' Imprint, 21*(15), 1–17.

Norcross, A. (1997). Comparing harms: Headaches and human lives. *Philosophy & Public Affairs, 26*(2), 135–167.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.

Parfit, D. (2017). *On what matters: Volume three* (Vol. Three). Oxford: Oxford University Press.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Talbot, B., Jenkins, R., & Purves, D. (2017). When robots should do the wrong thing, chapter 17. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 258–273). New York: Oxford University Press.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 6*(94), 1395–1415.

Thomson, J. J. (2008). *Normativity*. Chicago: Open Court.

Voorhoeve, A. (2014). How should we aggregate competing claims? *Ethics, 125*(1), 64–87.

Voorhoeve, A. (2017). Why one should count only claims with which one can sympathise. *Public Health Ethics, 10*(2), 148–156.

Williams, B. (1982). How to think sceptically about the bomb. *New Society, 62*, 288–290.