Training Essay

This is *not* the real essay question, but only an exemplary essay question for training purposes! This essay will *not* be corrected or graded by us.

General Remarks

The following remarks hold for the essay. Please make sure to follow them.

- You have to pass the essay in order to pass the course. You can get a bonus on your final mark for an especially good essay.
- Your essay needs to be your own work. While you are explicitly encouraged to discuss the issues with your fellow students and other people in general, group work is not allowed.
- You can write your essay in German or in English. However, decide for one of the two languages. If you write in German, you can still use the English philosophical terms from the lecture without translating them.
- Please leave a sufficiently large margin for corrections and comments. Be reasonable in your decision. Use a serif font, a font size of 11 or larger, and sufficient line spacing (≥ 1.5).
- Feel free to write what you personally believe to be correct. Don't try to give answers solely because you think that we would like you to give them. We will not grade you based on whether or not you align with our personal opinion. However – and this cannot be emphasized enough – there *are* right and wrong ways of reasoning, and we will grade you on that.
- If you use any kind of literature or online resources other than the lecture slides, make sure to cite them appropriately. Plagiarism will result in failing the course.

Training Essay

You have to write one coherent, appropriately structured text that presents an argument for a claim and thus tries to convince the reader of the argument. Your essay must have the following structure:

Introduction You should give short motivation why your topic is relevant, interesting, or otherwise worthy writing about.

Main argument The core part of your essay is your argument. This argument is to argue for or against the given claim or a reasonable conditionalization thereof. You should give a sound argument in extended standard form that is suitably embedded in your text. There should be sufficient soundness reasoning.

Most promising attack After having presented your argument, you should give the prima facie most promising attack of your argument. If you feel that there is more than one strongest attack, pick one of them. Describe the attack. State which aspect of your argument is attacked *and* explain, how the attack can be countered.

Conclusion In the end, briefly summarize the most important points you made.

Also, refer to the information given in the video on essays, and take note of the grading scheme.

Essay Topic – Autonomous Lying

As Bonnefon at al[1] showed, people would like others to use utilitarian cars, but would not buy one for themselves. A potential approach to this problem is to build cars that seem to make decisions that are in favour of the passengers most of the time, when in fact the car is utilitarian and just very good at making up excuses to let people think that they are usually deciding in favour of the passenger:

The claim: "We should implement cars such that they usually behave in a utilitarian way even though it is to the disadvantage of the passengers, and that they afterwards make up excuses to cover the fact that they behaved in a utilitarian way even though it was to the disadvantage of the passengers."

If you want to, you can abbreviate this claim with:
The claim (shortened): "We should build lying autonomous cars."

[1]Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. "The social dilemma of autonomous vehicles." Science 352.6293 (2016): 1573-1576.; available in the dcms

Saarland University
Summer Term 2020

Autonomous Lying

An Essay for the Lecture
"Ethics for Nerds"

Lecturers: Kevin Baum, Sarah Sterz, Holger Hermanns

Daniel Oster (251337)
Timo Speith (254242)
July 7, 2020

Contents

Notes on this essay

- This is an example of how an essay that solves the problem of the training essay could look. An essay like this would receive the bonus of +0.3 on the final grade.
- You can take this as an example of how a good essay can be structured. But please do not try to copy it contentwise.
  - Firstly, the argumentative ideas are not easily transferable to other topics and other claims.
  - Secondly, we usually want you to argue for what you personally believe to be correct, and not for what you think we want to hear. We do not grade you based on your opinion, but on how well your essay will be exerted methodologically. This also includes a reasonable amount of plausibility[1], but we usually make sure to pick claims for which, in the scope of an essay, reasonably plausible arguments for both sides can be found.
- Nothing specific in this example essay is mandatory to do, besides the things that are already specified in the problem statement of the essay. You can do things differently and still receive a bonus. In fact, for a different claim and a different topic, it most likely will be very helpful to do certain things differently.
- If your essay has similarities to this example essay in the end, you will not be graded down for plagiarism.

[1] Openly hateful positions, which include but are not limited to racism and sexism, have minimal plausibility.

1 Introduction

Autonomous cars can prevent a lot of harm by dramatically cutting down on accidents. Prima facie, most harm can be prevented if the car always behaved in a utilitarian way, minimizing the number of people affected by an accident, even if this would include sacrificing passengers for the sake of saving more people outside of the car.

But if we let the cars decide in the utilitarian way and let everyone know this, it is likely that only few or at least significantly less people will buy autonomous cars than in the case where we let them decide in favor of their passengers.[2] This, in turn, would seriously impede or even prevent the widespread introduction of autonomous cars.

One way in which this problem prima facie might be overcome is to implement cars such that they usually behave in a utilitarian way (even when this is to the disadvantage of the passengers), and that they afterwards make up excuses to cover the fact that they behaved in a utilitarian way. We call this the *autonomous lying approach* (shorthand: ALA). In this way, autonomous cars can find a widespread adoption while still saving as many lives as possible.

Our intuition tells us that there is something odd about ALA and that we are morally not allowed to take this approach. However, intuitions can fail. So, ALA may in fact turn out to be the morally best available approach, and if this was so, we should choose it. But we want to argue that our intuition in fact is accurate and well in line with a very broad range of moral theories.

In this essay, we argue that we should not build lying autonomous cars. First, we give the basic idea behind our argument, and then give our argument in extended standard form. Subsequently, we make this argument as plausible as possible by means of soundness reasoning. Finally, we discuss the most obvious point of attack and try to eliminate it.

[2] Cf. Bonnefon et al. 2016, "The Social Dilemma of Autonomous Vehicles". Notice that this is not exactly what the studies examined by Bonnefon et al. predict. The studies let us only estimate that significantly less people will buy a utilitarian car that sometimes kill its own passenger, than if the car always protects the passenger. But concerning the consideration here, we assume that we can extrapolate this result.

2 The argument

In this essay, we want to argue against the conclusion "We should build lying autonomous cars", which we take to be shorthand for "We should implement cars such that they usually behave in a utilitarian way even though it is to the disadvantage of the passengers, and that they afterwards make up excuses to cover the fact that they behaved in a utilitarian way even though it was to the disadvantage of the passengers."

2.1 The Basic Idea

If all three families of moral theories mostly agree on the deontic or normative status of an action, then this is very good evidence that the action in question really has this deontic or normative status. So, if we were able to show that all three families of moral theories by and large agree that pursuing ALA (for short: ALA-ing) is wrong or forbidden, we would need to assume that ALA-ing is in fact wrong or forbidden. If this was the case, then we should not build lying cars. This is the basic idea behind our argument. The general motivation behind lying cars is a consequentialist one. Considering this, it is not surprising that most deontological and virtue theories arguably judge ALA-ing to be wrong. We then are only left with consequentialist theories. There, we can distinguish between objective and subjective accounts. For the case we want to investigate, we need a decision procedure, and not only a criterion of rightness, because we want to know what we ought to do. Therefore, objective accounts can be disregarded. Hence, we are only left with subjective consequentialism. If subjective consequentialism judges ALA-ing to be wrong, we have to assume that all families of moral theories by and large agree that ALA-ing is wrong. Indeed, we think to have compelling reason that subjective consequentialism judges ALA-ing to be wrong. With this deliberations, we can state an argument against ALA in a tabular form.

2.2 The Main Argument

P1: We cannot judge what objective consequetialist theories would say about building lying autonomous cars.

P2: Subjective conseqeuentialist theories say that building lying autonomous cars under any plausible circumstances is forbidden.

P3: If we cannot judge what objective consequetialist theories would say about building lying autonomous cars, and subjective conseqeuentialist theories say that building lying autonomous cars under any plausible circumstances is forbidden, then we have to assume that building lying autonomous cars under any plausible circumstances is forbidden by consequentlialist theories.

C1: *Therefore*, we have to assume that building lying autonomous cars under any plausible circumstances is forbidden by consequentlialist theories. (P1–P3)

P4: Building lying autonomous cars under any plausible circumstances is against the spirit of deontological theories and virtue theories.

P5: If building lying autonomous cars under any plausible circumstances is against the spirit of deontological theories and virtue theories, then we have to assume that building lying autonomous cars under any plausible circumstances is forbidden by deontological theories, and virtue theories.

C2: *Therefore*, we have to assume that building lying autonomous cars under any plausible circumstances is forbidden by deontological theories, and virtue theories. (P4, P5)

C3: *Therefore*, we have to assume that building lying autonomous cars under any plausible circumstances is forbidden by consequentialist theories, and deontological theories, and virtue theories. (C1, C2)

P6: If we have to assume that build lying autonomous cars under any plausible circumstances is forbidden by consequentialist theories, and deontological theories, and virtue theories, we should not build lying autonomous cars.

C: *Therefore*, we should not build lying autonomous cars. (C3, P6)

2.3 Soundness reasoning

It is easy to see, that this argument is valid, as it has a very straight-forward structure in propositional logic. The more interesting question is whether the premises are true and, thus, the argument sound. In order to answer this question, we first provide a few considerations regarding the outcomes the introduction of autonomous cars and the use of ALA can have. Then, we will give reason for each of the premises.

As already mentioned, the seemingly best outcome (O1) that we can achieve, results from the successful introduction of autonomous cars and, in addition, from the utilitarian acting of the car in every decision situation. As also mentioned before, there are two other possible outcomes we can reach. One of them (O2) is generated by the failure of the introduction, the other (O3) by the success that results from cars deciding in favor of their passengers in dilemma situations (maybe more outcomes are possible).

However, if we take ALA, there is a viable risk that potential passengers, buyers, and finally the whole society will find out that the entire autonomous car industry is based on a deception, and that there is no favoritism of the passengers. If this is the case, it could create a general mistrust of new technologies that otherwise would have had good consequences and benefits but will never be introduced and used. Altogether, if we detect the deception, it could lead to a very bad outcome (O4). So, the possible outcomes connected to taking or not taking ALA are summarized in Table 1.

Outcomes:	O1		O3		O2		O4
Relation:		>		>		>	
Value:	best		2nd best		3rd best		worst
	(good)		(only slightly worse)		(a lot worse)		(really terrible)
'Action':	ALA-ing		not ALA-ing		not ALA-ing		ALA-ing

Table 1: The outcomes O1 and O4 are (according to our assumption) the only possible outcomes for the 'action' of taking ALA (ALA-ing) and the outcomes O2 and O3 are some possible outcomes for not taking ALA (not ALA-ing).

With this in mind, we will give reason for each of the premises.

P1: Objective consequentialism is not action guiding in this case, because we have a fair bit of uncertainty here, and we cannot be sure enough about the consequences of our action, let alone do we know the consequences of our actions beforehand. ALA-ing might bring about O1 or O4 – we just do not and cannot plausibly be sure enough in advance. So, we lack a crucial ingredient to tell what objective accounts would demand us to do, and therefore we cannot tell how they would judge ALA-ing.

P2: As seen above, we have to decide between three courses of action that can have four different outcomes. The first course of action is taking ALA (*ALA*-ing), the other is taking the passenger friendly approach (*PFA*-ing), and the last one is taking the faithful utilitarian approach (*FUA*-ing). We have to choose between three disjunctive actions, and we are in contemplation which one we ought to perform according to (subjective) consequentialism. Allegedly, ALA-ing has two possible outcomes, the best outcome (O1) and the worst outcome (O4). For the sake of the argument we will even grant that the worst outcome is very unlikely, as shown in Figure 1.

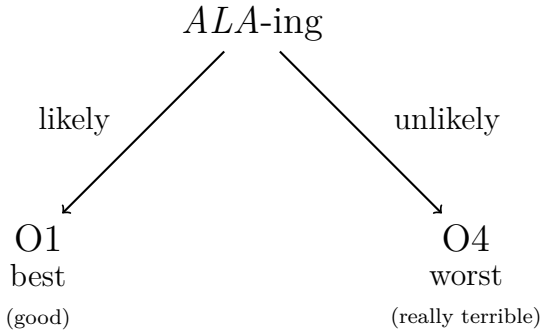

Figure 1: The possible outcomes of taking ALA

The setting is clear enough that we can gain some knowledge on the expected utility of ALA-ing. In the sense of subjective consequentialism, we take the likelihood of an outcome happening and multiply this likelihood with the value we would attach to this outcome to arrive at the value of the expected outcome. This means, we transform the outcomes O1 and O4 in one expected outcome O_{14}, as done in Table 2.

'Action':	PFA-ing		ALA-ing		FUA-ing
Outcomes:	O3		O_{14}		O2
Relation:		>		>	
Value:	best		2nd best		3rd best
	(good)		(somewhat worse)		(a lot worse)

Table 2: The expected outcomes.

As O3 is only slightly worse than O1, but O4 is really terrible, O_{14} is certainly worse than O3 but likely still better than O2.[3] With this we can see that ALA-ing is not the action with the best expected outcome, but PFA-ing. According to subjective consequentialism, we should take PFA and not ALA.

P3: We want to know in the end what we ought to do. So, we do not need a criterion of rightness (or, for that matter, not *only* a criterion of rightness), but a decision procedure. Objective consequentialism is not usable as a decision prodecure at all in this context, because we cannot tell how it would judge ALA-ing (see P1). Maybe objective consequentialism *does* say that it would be right to introduce lying autonomous cars. But we just have no way of knowing, and not even of somehow estimating in a reliable enough way. So, in order to decide what conseqentialism says about the matter, we have to rely on subjective consequentialism. If it says that pursuing the approach is forbidden, we have to assume that consequentialism as a decision procedure in general forbids it.

P4: The spirit of virtue theories, i.e. their core idea, is that it is right to be virtuous or to behave virtuously, and wrong to be vicious or behave viciously. Deceiving people that trust you on a large scale prima facie is clearly not virtuous, but vicious. The core idea of deontological theories is that it is wrong to violate certain universializable rules. Deceiving people on a large scale arguable is not in line with most or all plausible

[3] Whether O_{14} is better or worse than O2 actually does not matter for our argument, as long as O3 is still the best outcome.

candidates for these rules. Therefore, ALA-ing is against the spirit of both deontological and virtue theories.

P5: If something is against the spirit of the theory, i.e. the core idea of this theory, then we have to assume that this theory forbids it. This is a very plausible claim. Afterall, why would a theory be in favor of something that violates its very core idea? (Note the hedging here: the premise is not a metaphysical claim about how the world is, but an epistemic claim about what we have to assume the world to be.)

P6: The last premise just has a great intuitive pull, at least if we are talking about decision procedures – which we are. The three families of moral theories are very diverse and all have their strengths and weaknesses. They cover the space of intuitively plausible moral theories, and there are no hot candidates besides them. So, if we subscribe to any of those, we have to assume that ALA-ing is wrong, and therefore that we should not pursue ALA.

2.4 Attack

One of the premises that might be attacked easiest is P5. After all, there could be theories of the one or the other family that are in favor of pursuing the approach. First, the talk of the spirit of a theory is very imprecise and nebulous. Second, even if we grant that something like this exists and can be spelled out in a suitable way, it is still the case that not every theory has to be in line with the alleged spirit of a family of moral theories in all cases. There might be corner cases where a plausible theory goes against the spirit of its family, or where it fulfills the spirit in a prima facie unexpected way. An attack might then go like this: it could be that there are specific deontological or virtue theories, that are in favor of ALA-ing. P4 is insufficient to disprove this, and if it turned out that there actually are theories like this, ALA-ing would then not be universally disallowed by the two families. Then, C2 and C3 would not follow anymore, and ultimately the main argument would be unsuitable to support C.

2.5 Defense against the attack

Many examples of plausible moral theories from the two families *generally* speak against such approaches as ALA:

First, let us have a look at deontological theories. Every account that is breathed on by a Kantian touch will never allow lying and, consequently, it will not support ALA. Not lying is a perfect duty according to Kant (as seen in the lecture) and, consequently, lying is always forbidden. However, Kant is not the only proponent of deontological theories. We have seen in the lecture that, prima facie, Scanlon would be in favor of a lottery in cases where lives are involved. Therefore, according to Scanlon's contractualism, we are not allowed to introduce utilitarian autonomous cars. Consequently, we are also not allowed to introduce lying utilitarian cars. Let us give a final example in the realm of deontology. People in the original position would probably want to use utilitarian cars openly (FUA), and would therefore not be in favor of ALA. This is so because they would follow a minimax strategy, and they do not know whom them will become, so they would ceteris paribus always save the larger group and thereby maximize their own chances of living.

Second, from a virtue ethical stand point, we can take a look at the two example theories from the lecture: Eudamonism and Exemplarism. According to Eudamonism, a person is not allowed to do what a perfectly virtuous version of this person would not do. It is very plausible that ALA-ing is a thing a perfectly virtuous version of most people would not do. Deceiving others is, as such, rather a vice than a virtue. For these reasons, it is unlikely that ALA-ing constitutes a case where deception is done by a virtuous person. Exemplarism takes a slightly different approach to Eudamonism. We do not look at what a virtuous person would do, but at whether an action expresses first and foremost a virtue or a vice. Since ALA-ing is inherently connected with deceiving others, we can assume that the action expresses a vice, not a virtue. Thus, Exemplarism also speaks against ALA.

This paints a pretty clear picture already. We have now seen that the core idea of both families is against ALA-ing, *and* that many important proponents of the two families are in line with this core idea. So, the response

to the above objection is: Yes, it could be that there are plausible theories in favor of ALA-ing, but the burden of proof lies on the site of the person who claims that there is such a theory. If objectors of our argument can point to such a theory, our argument has to be revised. But until then, it is plausible and suitable to support its conclusion.

3 Conclusion

We argued for the claim that we should not take the autonomous lying approach under all plausible circumstances. Out of the perspective of the three major moral theories – especially out of a (subjective) consequentialist perspective – we should not take ALA. We gave an argument for our claim and defended this argument against its strongest attack. The burden of proof that we should ALA now lies on the side of our opponents.

References

Bonnefon et al 2016. Jean-François Bonnefon, Azim Shariff and Iyad Rahwan: „The social dilemma of autonomous vehicles“, *Science*, 23 June 2016, 352 (6293), 1573-1576, URL = http://science.sciencemag.org/content/352/6293/1573.