
Dependable Systems

dcms.cs.uni-saarland.de/ethics_21/
Sarah Sterz, Kevin Baum, Prof. Holger Hermanns

Training Exercises C3 (Manipulation, Deception, and Illusion)

with Example Solutions

Issue 1: *#varoufake*

Jan Böhmermann created a lot of confusion with Varoufake. Was he morally allowed to claim that he faked the video? Was he disallowed to do so? Give your intuition and reason to believe in your intuition.

If you want to see more check out the original clip from the show: `https://youtu.be/Vx-1LQu6mAE` (English subtitles are available)

Issue 2: *Ought I to publish?*

Think about the following claims and their implications. Do you think that they are true?

1. Researchers are morally forbidden to publish research results that can have very drastic negative consequences.
2. That people loose trust in media is a very drastic negative consequence.
3. Open-source tools that allow for easy and convincing video manipulation can have as a consequence that people loose trust in media.
4. Deep fake should never have been published.
5. There are no countermeasures to the negative consequences of apps like Deep Fake.

Issue 3: *Red Flag Laws*

Make a moral case for or against the following claim or a *reasonable* conditionalization thereof:

> Introducing a Red Flag Law for AI is morally right.

Give a compelling argument in extended standard form that can, but does not have to, incorporate the use of moral theories. Make sure to give sufficient reason for your premises.

Sketch of a Solution 3:

Argument:

P1:	It usually will not lead to any harm or reduces benefits when an AI communicates that it is an AI.
P2:	It can lead to harm and usually does not increase benefits when an AI does not communicate that it is an AI.
P3:	If P1 and P2, then introducing a Red Flag Law for AI has a higher utility than not doing so.
C1:	Therefore: Introducing a Red Flag Law for AI has a higher utility than not doing so. (P1–P3)
P4:	If introducing a Red Flag Law for AI has a higher utility than not doing so, then C.
C:	Therefore: Introducing a Red Flag Law for AI is morally right. (C1, P4)

In most circumstances in which AI would be used in direct social interaction with humans, like service or entrainment, the AI will plausibly not perform any worse in its task if it is made transparent to the humans (and maybe even to other AIs) that they are interacting with an AI. On the other hand, if a human mistakes an AI for another human, this might lead to dangerous situations, e.g., when the human is in an emergency situation and requires help that the AI is not equipped to give. It plausibly would be better for the human if they knew that they were interacting with an AI and could directly seek help elsewhere.

P4 is correct if we employ consequentialism. (This argument does not reflect other moral theories, but plausibly they would agree. It can, e.g., be argued that undisclosing AIs are deceptive, which would probably not be favoured by Kant. Likewise, it is plausible to think that someone could reasonably reject any set of principles that allowed for not introducing a Red Flag Law in Scanlonian Reasoning.)

Issue 4: *Beauty Filters*

Familiarize yourself with the effect that the exposure to retouched photos can have on users, especially on teenagers. You might want to take a look at what is sometimes called "Snapchat Dysmorphia". Try to answer the following questions:

(a) Is it morally permitted to use beauty filters for a given user?

(b) Is it morally permitted for online services to promote the use of beauty filters?

(c) What does this have to do with the collective action problem?

If you want to know more about the impact of social media on the wellbeing of teenagers and young adults, check out 2017's #StatusOfMind report: `https://www.rsph.org.uk/our-work/campaigns/status-of-mind.html`

Sketch of a Solution 4:

(a) Notes: It is not asked for an argument here (but we gave one just for the sake of it).

Argument:

P1:	If someone uses a beauty filter on a particular photo, it usually has no or only very small negative consequences.
P2:	If someone uses a beauty filter on a particular photo, it usually has small positive consequences.
P3:	If P1 and P2, then it is permitted to use a beauty filter on a particular photo.
P4:	If, according to consequentialism, it is permitted to use beauty filters for a given user, then it is morally permitted to use beauty filters for a given user.
C:	It is morally permitted to use beauty filters for a given user.

P1 is the case, because there is usually no harm in just altering one particular photo. (What can be harmful is if altering pictures becomes a habit, but it is hard to argue that a single filtered photo is already harmful.) P2 holds because it might be fun to alter an image, make oneself feel more attractive, give oneself a shot-term boost in confidence or increase one's popularity.

(b)

Argument:

P1:	If beauty filters get promoted, then we have to expect that many people will use them.
P2:	If we have to expect that many people will use beauty filters, then we have to expect that an unrealistic beauty standard.
P3:	If we have to expect an unrealistic beauty standard, then we have to expect that this will have an overall negative effect on people's well-being and mental health.
P4:	Teenagers developing inferiority complexes and so on is a bad consequence.
C1:	Therefore, if beauty filters get promoted, then we have to expect that this will have an overall negative effect on people's well-being and mental health. (P1–P4)
P5:	If C1, then it is not morally permitted for online services to promote the use of beauty filters.
C:	It is not morally permitted for online services to promote the use of beauty filters. (C1, P5)

(c) If few people use beauty filters or people only use them very sparingly, the impact on society probably is rather small. If, however, many people use beauty filters to a rather large extent, this likely leads to unrealistic standards of beauty. This may cause all kinds of problems, like teenagers having inferiority complexes and so on. Using beauty filters can be seen as a kind of collective action problem: as altered photo more or less does not change someone's standards of beauty, let alone a whole society. So, altering a particular image is permitted. But if everybody does so, this can have very bad consequences. If nobody used beauty filters, however, these bad consequences would not come about. So, this is a typical case where one person probably does not have any impact, but the collective action of many has.

Issue 5: *Free Choice*

(a) If you were to explain the either-or problem of freedom to your mother, how would you do it?

(b) Write down the argument from the either-or problem or freedom as an extended standard form.

(c) Do you find the conclusion of the problem scary?

(d) Do you agree with the argument or do you have an objection in mind?

Sketch of a Solution 5:

(a) Hi mom, I learned something shocking today in my ethics lecture: Actually, free will might not exist. There are basically two ways to think of the future: either as something that is predetermined (by laws of physics, or some heavenly creature or whatever), so there is just one possible future, or the future is not determined, so there are different possible futures. This is also true for each individual action, either it is determined or not. If it is determined, it obviously cannot be changed by our free will. But if it is not determined, it is pretty much a random act. And if an act is random, it is also not guided by our free will. So, either way, there is no free will!

Long story short: Sorry for not emptying the dishwasher, it just was determined this way or didn't occur as a random act! I had no chance to do it. ;)

(b)

Argument:

P1:	All acts are determined or not determined.
P2:	If an act is determined, then is it unfree.
P3:	If an act is not determined, then it is random.
P4:	If an act is random, then it is unfree.
C1:	Therefore: If an act is determined or not determined, then it is unfree. (P2–P4)
C:	Therefore: All acts are unfree. (C1, P1)

(c) Do you find the conclusion of the problem scary?

(d) Do you agree with the argument or do you have an objection in mind?

Issue 6: *Dark Patterns*

During the video, it was said that Dark Patterns usually are bad and that employing them ceteris paribus is not morally permissible. Why is this true?

Issue 7: *Nudging*

In the lecture we very briefly discussed whether it is ok (rationally and morally) to take someone's autonomy away with nudging. Now look at the three aspects of this that probably are most relevant for your private and professional life. Argue for or against the following claims or a reasonable conditionalization thereof:

(A) It is rational (i.e. in my best interest) to avoid being nudged by technologies.

(B) It is morally obligatory to avoid being nudged by technologies.

(C) It is morally permissible to implement a nudging technology.

During this process, also think about the following questions:

(i) Where do you use nudging technologies?

(ii) Do you try to avoid nudging and if yes, where and why?

(iii) Do you think it makes a difference whether you purposefully decide to use a nudging technology, or whether you use it without realizing that is (or tries to be) nudging?

Sketch of a Solution 7:

Argument: (A)

P1:	Being nudged through technology can take away one's autonomy.
P2:	If I do not deliberately decide on having my autonomy taken away in advance for a very good reason, it is rational to retain one's autonomy.
P3:	If P1 and P2, then C.
C:	Therefore, if I do not deliberately decide on having my autonomy taken away in advance for a very good reason, it is rational to avoid being nudged by technologies.

Argument: (B)

P1:	If it is morally obligatory to avoid being nudged by technologies, then I am doing something wrong each time I deliberately and knowingly use a nudging technology in order to be nudged by it.
P2:	If I am doing something wrong each time I deliberately and knowingly use a nudging technology in order to be nudged by it, then using a well-designed fitness tracker to be more active is always wrong.
P3:	Using a well-designed fitness tracker to be more active is not always wrong.
C:	Therefore, it is not morally obligatory to avoid being nudged by technologies.

Argument: (C)

P1:	If nudging technologies help people to adopt behaviours that are beneficial for them and the society, and the use of which has little or no negative consequences, then these nudging technologies have an overall positive influence on people's wellbeing.
P2:	If nudging technologies that help people to adopt behaviours that are beneficial for them and the society have an overall positive influence on people's wellbeing, then C.
C:	Therefore, it is morally permissible to implement a nudging technology that help people to adopt behaviours that are beneficial for them and the society, and the use of which has little or no negative consequences.

For (B) and (C), think for example of fitness trackers. It is both allowed to implement fitness trackers (if they are designed well and do not, say, have privacy issues) and to use them and deliberately be nudged by them.

Issue 8: ***Electronic Whip***

Research on what is called Disneyland's "electronic whip". Where does this belong in the realm of what we discussed in the module? Does any of this sound familiar to you? In fact, the Tim-case from the Ethics exercises was inspired by this. Where are similarities and where are differences between the two cases? How do you assess the moral status of implementing either technology?

Issue 9: ***Autonomy***

Which of the definitions of autonomy do you find most compelling and why? Can you come up with a counterexample against one of them?