We Are AI: Taking control of technology

We are AI #3: Who lives, who dies, who decides?

Cover-alt

A magnificent golden can stands opened – the lid of the can is opened upwards, and we see the classic trolley problem scene in its reflection: a tram/trolley approaches a track that is diverging into two paths. On one path lies one person tied to the track, while on the other path is a group of 4 people tied to the track. A man stands at the intersection, controlling a lever that decides which track the trolley goes on. The can is overflowing with pink worms, some of which have fallen onto the floor. The worms have abstract patterns on them and the words “inequality”, “values”, “utilitarianism”, “uncertainty” and “ethics” written on them. It is symbolic of the idea that the trolley problem opens a can of worms of ethical issues in AI.

Cite as:
Julia Stoyanovich, Mona Sloane and Falaah Arif Khan.
“Who lives, who dies, who decides?”. We are AI Comics, Vol 3 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol3_en.pdf

Page 1

Prediction is difficult, especially of the future.

Difficult as it is - because of the uncertainty and complexity of the world - predicting the future is often the job of AI.

A fortune-telling AI sits in front of a crystal ball, with its hands waving over the ball, its expression severe, and eyebrows raised.

And because this task is difficult - and at times even impossible - AI systems will make mistakes.

For example, a smart light AI may incorrectly guess whether the light should be on or off.
Turning on the light in the middle of the night will wake you up.

The sleeping woman is abruptly woken up by blinding lights– zombie-like eyes and disheveled hair mirror the exasperated expression on her face and her arms thrown in the air.
And leaving it on systematically will leave you to foot a high energy bill.

A woman holds a massively long paper of the electricity bill in her hand, and exclaims upon looking at the estimate. Her hand is on her head in astonishment, and she looks distressed.
As another example, a customer service AI at your favorite shoe store may misunderstand your order, …and the wrong pair of shoes will be shipped to you.

We see a conversation between a customer care AI and a customer. The customer care AI is wearing a headset and speaking into the microphone, and a woman is holding up a pair of vehemently colored and patterned shoes, while frowning.

Annoying as they may be, these are mistakes with low stakes.

Consequences of such mistakes are not severe, and they are reversible.

Page 2

However, there are cases where mistakes can lead to catastrophic irreversible harms,
even to the loss of human life.

Consider an autonomous car -

A woman kicks back in an autonomous vehicle – she is laying back in the driver seat with her legs on the dashboard, as the steering wheel lights up and the car goes into self-drive mode.
an AI that is about to cross an intersection,

and that does not recognize a person on a bicycle as one of the types of objects it would expect to see on the road.

The car would then continue on its path, running the cyclist over.

We see the driver's view of the self-driving car. There are two cars on the street across, trees on the sidewalk, and a female bicyclist crossing in front. An image segmentation algorithm is identifying different objects in the scene – it has detected the tree and the cars and flags them up with positional boxes. It has failed to notice the bicyclist as an object on the road.
Another example is when the autonomous car does not detect the presence of a person in a wheelchair crossing the intersection.

This could happen if, for example, the person were crossing the intersection going backwards,
and the self-driving car's AI miscalculates the pedestrian's trajectory.

A woman on a wheelchair is crossing an intersection. The traffic light is red, and the woman is turning her wheelchair around and crossing backwards to the direction she is facing.
But human drivers also cause accidents!

So why let perfect be the enemy of good?

A caricatured Elon Musk shrugs and points to the right of the screen, making the case for self-driving cars.

Shouldn't we be prepared to suffer a few mistakes made by autonomous cars in the name of increased overall safety of our transportation system, and the convenience to the drivers?

Page 3

In fact, can't we encode our judgment about what mistakes are more important to avoid, and let an AI sort out the trade-offs?

Can't we equip our AI with values?

A famous example that makes us think about our values, and trade-offs they introduce, is
The Trolley Problem.

It is a thought experiment that raises an ethical dilemma:
Should we sacrifice the life of one person to save the lives of a large group of people?
Interestingly, experiments in ethics and psychology have shown that there is no clear-cut answer.

What we decide depends on our values - on what we consider right or wrong, on the various elements of our identity, on our cultural background, and also on the specific set-up of the problem: on the context in which the decision is being made.

Picture a massive, full-page long network of trolleys and intersecting paths –
From top to bottom: a trolley approaches a track that is diverging into two paths. On the upper path lies one person tied to the track, while on the lower path is a group of 4 people tied to the track. A man stands at the intersection, controlling a lever that decides which track the trolley goes on. The lower path then leads into another diverging path, with the similar upper and lower paths with a single person and a group of people tied to it. Each path keeps diverging into subsequently more paths, showing the interconnected nature of these "trolley problems", and the compounding impact of individual decisions.

Interesting as it is, the trolley problem is still a thought experiment, and it has been criticized as being so outrageous as to be unrealistic.

Page 4

But self-driving cars are now presenting us with a real-world version of this dilemma.
If we decide to broadly deploy self-driving cars, then how do we deal with the mistakes that are bound to happen, even if there are relatively few such mistakes?

… and what about an entire transportation system made up of autonomous cars, people, weather, and different road conditions -
How do we simultaneously deal with hundreds of mutually-dependent trolley problems?

An important additional difficulty is that, in contrast to the classic trolley problem, where it is known how many people are on what side of the track, an autonomous car — and other types of technology — operate under a high degree of uncertainty.

It may be unknown whether there are even people on the tracks, let alone how many of them there are, and which groups they may represent.

How do we make value judgments in the face of uncertainty?

Picture a variation of the classic trolley problem under incomplete information — A trolley approaches a diverging path. A man stands at the intersection controlling a lever that decides which path the trolley will take. The man suspects that there are people tied to the tracks, and thereby human lives will be lost if the trolley continues on, but he is unsure about how many people, who they are, etc.

Page 5

The trolley car problem illustrates a specific doctrine of moral philosophy - Utilitarianism
Perhaps this doctrine can offer us some guidance?

Utilitarianism is a moral principle that holds that the right course of action — in any situation —
is the one that produces the greatest balance of benefits over harms for everyone affected.
Utilitarianism stems from the late 18th- and 19th-century English philosophers and economists Jeremy Bentham and John Stuart Mill.

A caricatured Jeremy Bentham stands and poses at the camera/viewer.

Bentham famously said: “It is the greatest happiness of the greatest number that is the measure of right and wrong.”

Sounds great indeed!

An embodied AI/robotic figure “showers” people with happiness – a large group of people are dancing/jumping in celebration as a massive robot stands over them and sprinkles particles of the “smiley face” emoji and the “love” emoji, in the style of internet sensation Salt Bae.
Unfortunately, applying these ideas to self-driving cars — and to the design and operation of technology more generally — opens a can of worms.

And it has a name:
Algorithmic Morality
An opened can lies on its side, with its contents falling out. The lid of the can is open to the right, and the reflection in it shows the trolley problem. The contents of the can are little worms that are falling out the can.

Page 6

Algorithmic morality is the act of attributing moral reasoning to algorithms.

Two women stand on either shoulder of an embodied AI/robot. The one on the left is dressed like a devil – with a pitchfork in hand, horns on her head, and dressed in red from head to toe. The one on the right is dressed like an angel – with a halo and wings, dressed in white from head to toe. Both women are gesturing and speaking to the AI. The AI looks perplexed – with it’’s eyebrows raised and its teeth clenched.

Doing so is problematic. Here is why.

To start, how do we measure happiness and unhappiness?

A woman smiles widely at the camera/viewer. A pair of hands are using a measuring tape to determine the width of her smile, as a proxy for the amount of happiness.

And how do we then encode these measurements into a set of objectives that an algorithm will understand?

There rarely exists a mathematical formula or a logical statement that can capture the balance between the benefits and the harms.

A scientist is frantically writing formulas on a whiteboard. She is spectacled and is holding a marker pen, with which she scribbles formulae for happiness, sadness, amount of benefit and degree of harm.

In other words: there simply isn't a formula for "right" or "wrong".

And there isn’t a formula for values, and for how values emerge and change in complex social situations.

Another reason why algorithmic morality is problematic is that, when a mistake in judgment about what is right or wrong is made — and, as we already know, mistakes will be made because the world is complex, uncertain, and perhaps even unpredictable —
algorithmic morality would require an algorithm to take responsibility for the mistake.

An embodied AI/robot’s mugshots are being taken – we see it in two different poses and holding a black board: one facing the viewer/camera, and the other facing sideways guiltily looking at the viewer.

Page 7

But holding an algorithm responsible for a mistake makes no sense:

An embodied AI/robot is puppeteering a human – it holds a human puppet in its left hand and is pulling on strings which makes the human wave its arm and look up at it, while its right arm is folded onto its hip.

an algorithm does not possess consciousness or free will, it does not make an intentional choice that leads to a mistake, and so cannot be held accountable.

Where does this leave us?

The can-opener that is the trolley problem showed us that we cannot delegate ethics to machines.

That it is still up to us, humans, to make choices and take actions (or choose not to act),
in accordance with our values, and with existing laws.

And then it's up to us to take responsibility for the consequences of any mistakes.

A human is smiling while holding a robot-puppet – she's pulling on strings to make the robot do a little dance.

We cannot outsource the work of being human to a machine.

Page 8

In summary, to embed ethics into socio-technical systems such as AI, we must think about what values are baked into these systems, who benefits when the systems work well, and who is harmed by their mistakes.

And we must collectively take responsibility for deciding on the balance between the benefits and the harms, so that "the greatest happiness" that Jeremy Bentham promises to the greatest number of people is also enjoyed by the greatest diversity of stakeholders.

This work of collectively understanding and negotiating the trade-offs is what roots the design of technology in people.

Picture a large “AI-tree”: the leaves are composed of a large network/graph – with colorful and densely interconnected nodes and edges. The roots of the tree span a large area underground, where a group of humans are holding on to the roots, and “grounding” the tree. The humans are drawn in abstract line art, and are only identifiable by a silhouette. They are in a variety of colors.