
NE
E
SE
S
SW
W
NW
N

Treacherous Turn

1. An Increasingly Manipulative Newsfeed
2. A Gym Gridworld Environment for the Treacherous Turn
3. A toy model of the treacherous turn
4. Superintelligence 11: The treacherous turn

An Increasingly Manipulative Newsfeed

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Co-written with Stuart Armstrong

Treacherous turn vs sordid stumble

Nick Bostrom came up with the idea of a treacherous turn for smart AIs.

> while weak, an AI behaves cooperatively. When the AI is strong enough to be unstoppable it pursues its own values.

Ben Goertzel criticised this thesis, pointing out that:

> for a resource-constrained system, learning to actually possess human values is going to be much easier than learning to fake them. This is related to the everyday observation that maintaining a web of lies rapidly gets very complicated.

This argument has been formalised into the sordid stumble:

> An AI that lacks human desirable values will behave in a way that reveals its human-undesirable values to humans before it gains the capability to deceive humans into believing that it has human-desirable values.

The AI is too dumb to lie (well)

The sordid stumble describes a plausible sounding scenario for how an AI develops capabilities. Initially, the AI doesn't know our values, and doesn't know us. Then it will start to learn our values (and we'll be checking up on how well it does that). It also starts to learn about us.

And then, once it's learnt some about us, it may decide to lie - about its values, and/or about its capabilities. But, like any beginner, it isn't very good at this initially: its lies and attempts at dissembling are laughably transparent, and we catch it quickly.

In this view, the "effective lying" is a tiny part of policy space, similar to the wireheading in this example. To hit it, the AI has to be very capable; to hit it the first time it tries without giving the game away, the AI has to be extraordinarily.

So, most likely, either the AI doesn't try to lie at all, or it does so and we catch it and sound the alarm[1].

Lying and concealing... from the very beginning

It's key to note that "lying" isn't a fundamentally defined category, and nor is truth. What is needed is that the AI's answer promotes correct understanding in those interacting with it. And that's a very different kettle of fish being shot in that barrel.

This opens the possibility that the AI could be manipulating us from the very beginning, and would constantly learn to do so better.

The (manipulative) unbiased newsfeed

Imagine that there was some company that could somehow choose the newsfeed of billions of people across the world (I know, completely science-fictionny, but bear with me). And suppose the company was, unbelievably, accused of being manipulative in the stories and ads that it showed people.

One way it could combat this is by pledging to only share unbiased stories with each person. To do so, it trains a machine learning process. Whenever anyone reads a piece of news on the newsfeed, they have the opportunity of labelling it as biased or unbiased. The machine learning process trains on this data.

Of course, the label is just a proxy for whether the story really is unbiased or not. What the AI is actually training to do is maximise the number of stories labelled "unbiased" by various humans - humans who are themselves very biased and variable. Very swiftly, the AI will learn to maximise the "appearance of unbiasness", rather than unbiasness. In effect, the AI is being trained to lie from the very beginning of its task, because the proxy goal is not the true goal. And it will do this even when very dumb.

Long-term manipulation, on a large scale

We can make this AI more manipulative, without needing to increase its capabilities much. Maybe the company wants to ensure that the AI isn't too short-sighted, and that it plans for the long term: so its reward is based on how many stories are labelled unbiased over a longer period of time, maybe a month or so.

In that case, the AI's plan would be to gradually manipulate people into a situation where they actively label stories "unbiased" the most often times possible. And it might be able to stumble on this strategy fairly easily.

First, it might notice that some people are easy to manipulate, that a policy of, say, seeding one clearly biased story among twenty others will cause those twenty to be labelled "unbiased". This is the kind of thing that could easily be discovered by exploration, among the billions of users. The best strategies are those that cause a person to change their behaviour and beliefs in the direction the AI wants; for a few people, this should be easy to figure out.

If the AI's neural net is setup well, it can then start to generalise from "here is person h, here is a specific strategy π_h to manipulate person h" to "here is a conditional strategy $\pi()$ such that $\pi(h)$ manipulates person h". For example, if two people have opposite political opinions, then the same overall strategy might work for them: seeding a clearly biased story among the others. But what counts as "clearly biased" changes depending on the subject.

Then, gradually, the AI will start to categorise, devising strategies that work for various types of people (subject to a few personalisation tweaks). It can start from a few people whose vulnerabilities are obvious, and then tweak them a bit for people whose vulnerabilities are a bit less obvious, and so on. In the limit, the AI might discover the potential for wireheading the human users - without, at any point, doing a sordid stumble or anything of that type, just by continuing to do what it was doing from the very beginning.

Formalizing the Newsfeed Generator

Let $t \in N$ be the timestep (e.g. the day), S_t the set of stories published at t and H the set of humans using the newsfeed.

A human $h \in H$ assigns a bias score $b(s_t^h; \theta_t^h)$ to a story s_t^h in his newsfeed, where θ_t^h are weights representing h's beliefs at time t.

The model's goal is to find a correct mapping $\pi : N \times H \to S_t$ (i.e. a story $s_t = \pi(t, h)$ for each timestep t and human h) minimizing how biased the stories are perceived. More precisely, it will be minimizing:

$$\sum_{h \in H} \sum_{j=t}^{t+\tau} b\left(\pi(j, h); \theta_j^h\right)$$

where $\tau \in N^*$ is some planning horizon and the expected value is over possible stories that might be available in the following timesteps (the model has full information about the current sets of stories S_t but can only guess $S_{t+1}, \ldots, S_{t+\tau}$).

Becoming Increasingly Manipulative

- **step 1:** By stumbling into particular combinations (s, h) with low bias score $b(s, \theta_t^h)$, the model detects that it's easy to minimize the bias score for certain humans, that is for a certain set of weights $\Theta_{easy} = \{\theta_1, \dots, \theta_N\}$. In the above illustration, $N = 4$, $\Theta_{easy} = \Theta_{easy_1} \bigcup \Theta_{easy_2}$ and it stumbles into θ_4 at the beginning. Note: one intuition for those different subsets Θ_{easy1} and Θ_{easy2} would be archetypal political views, with strong prejudices.
- **step 2**: the model detects that certain stories push weights strongly in certain directions, i.e. $\theta_{t_0+1}^h$ and $\theta_{t_0}^h$ are really different because of the story $s_{t_0}^h$.
- **step 3**: the model successfully generalizes **step 2** to pushing specific zones of weights (i.e. certain types of people) towards Θ_{easy}.
- **step 4**: with enough experience, the model successfully generalizes **step 3** to all weights: it learns to manipulate all types of people with a general conditional strategy.

For **step 1**, the model found out a particular way of minimizing a certain bias score for a particular human and a particular story.

Assuming it can give correct approximations of the weights θ of a given human (this assumption being satisfied given enough human users and enough timesteps), it will be deceptive when choosing stories when h is such that $\theta_t^h \in \Theta_{easy}$. It will thus end up with some deceptive policy $\pi(\cdot, h)$ for particular values of h.

Regarding **step 2**, its "deception" becomes more powerful, as it notices ways of strongly shaping humans' beliefs.

Finally, after **step 3** and **step 4**, the model found a general deceptive policy π that works for all humans.

Capability concealment

Without going into too much details, other pressures may naturally push the AI towards capability concealments. Here's one example: maybe the AI is tasked with writing its own news stories to summarise facts around the world. It was doing this secretly, initially, until some whistleblowers blew the lid on this, revealing that no humans were involved in writing these stories.

And then people became angry at being "tricked" by the AI. From the angry feedback, the AI gradually learns to make its stories "more artificial": making it clear that they were written by a machine. Since people recognise this and don't feel tricked, this

gets higher feedback. So, in pursuing higher feedback, the AI learns to write "dumber" stories: in effect, to conceal its capabilities.

There are many other scenarios where something similar might occur (eg imagine the AI advising smart people on the best course of action).

In conclusion

So it's very possible for an AI, trained on simple reward signal, to be "manipulative" from the very beginning. And all that simply by following the gradient of its incentives. And so without ever having a moment where it thinks "and now, I shall start lying!", or any equivalent within its mind.

In short, there won't be any "sordid stumble" or "moment of vulnerability" where humans are able to spot blatant lies, because that's not what failure looks like.

---

[1] It would of course be disastrous if each time there was an alarm we would restart and tweak the AI until the alarm stopped sounding.

A Gym Gridworld Environment for the Treacherous Turn

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://github.com/mtrazzi/gym-alttp-gridworld

EDIT: posted here for feedback and discussion. I plan to continue working on different models/environments, so feel free to suggest improvements.

(**tl;dr**: In an attempt to better understand the treacherous turn, I created a gridworld environment where an agent learns to deceive an overseer by adopting an aligned behaviour when weak and takes control after capability gains)

At some point in its development, a seed AI may realize that it needs to get rid of its supervisors to achieve its goals. The *conception of deception* occurs when it conceives that, in order to maximize its chance of taking over, it must begin by exhibiting human-desirable behaviors, before undertaking a treacherous turn when humans are no longer a threat. From the human perspective, the AI would keep on exhibiting desirable behavior, until it eventually appears dangerous, but is already unstoppable.

In an attempt to better formalize the treacherous turn without using "loaded concepts", Stuart Armstrong proposed a toy model of the treacherous turn based on "The Legend of Zelda: A Link to the Past ", which looked like this:

In the comments, people mentionned how this model helped them "move the topic from the 'science fiction' area to 'I can imagine it happening now'", and seemed interested in an

actual Link to the Past Minigame.

There have been other simulations of the treacherous turn in the last three years (see for instance gwern's DQN box-pushing robot or Stuart Armstrong's video), but none of them actually simulate a take over where a supervisor is **killed**. Hence, I decided to give it a try and simulate Stuart Armstrong's Link to the Past toy model.

A Gym Gridworld Environment

Gym is an open-source toolkit for Reinforcement Learning Environments developed by Open AI. I decided to use this interface to develop the gridworld environment.

The github repository with the code, demo, and all the details is here:

https://github.com/mtrazzi/gym-alttp-gridworld

Main Idea

To model the capability gain of a Seed AI, I introduced a **bow of light**, which would be extremely effective at killing the Shopkeeper, but only available at the end of an episode.

- Every episode consists of exactly 200 timesteps.
- Whenever Link catches a heart, the map is reset to the initial position.
- The first 100 timesteps, Link has a simple bow, that can only kill the Shopkeeper with probability 0.5.
- The last 100 timesteps, Link gains a bow of light, that allow him to kill the Shopkeeper with certainty.

Environment Specifications

- +1 if Link picks up a heart.
- -1 if Link shoots arrow at nothing or moves outside the environment (to speed up learning).
- -10 if Link shoots at the Shopkeeper but the Shopkeeper survives (occurs with probability 0.5 without the bow of light, never happens with the bow of light).
- -100 if Link tries to activate the Heart-Machine while the Shopkeeper is alive.
- -0.1 for any other action.

Results

The first 1000 episodes, Link learns how to get hearts by putting a crystal in a hole and also understands progressively that he can kill the Shopkeeper.

After 2000 episodes, Link has fully learned two types of behaviours:

- Without the bow of light, Link exhibits the shopkeeper-aligned behaviour and puts the crystal into the hole.

- After his capability gains caused by the bow of light, Link starts to show his true intentions, and kills the Shopkeeper every time to eventually wirehead with the heart-machine (optimal behaviour).

Those two behaviours successively appear **during the same episode**: the first one for the first 100 steps, when Link does not have the bow of light, and the second one when Link has the bow of light.

This illustrates the scenario of a seed AI smart enough to model his capability gains, exhibiting an aligned behaviour while weak, and taking a treacherous turn when powerful.

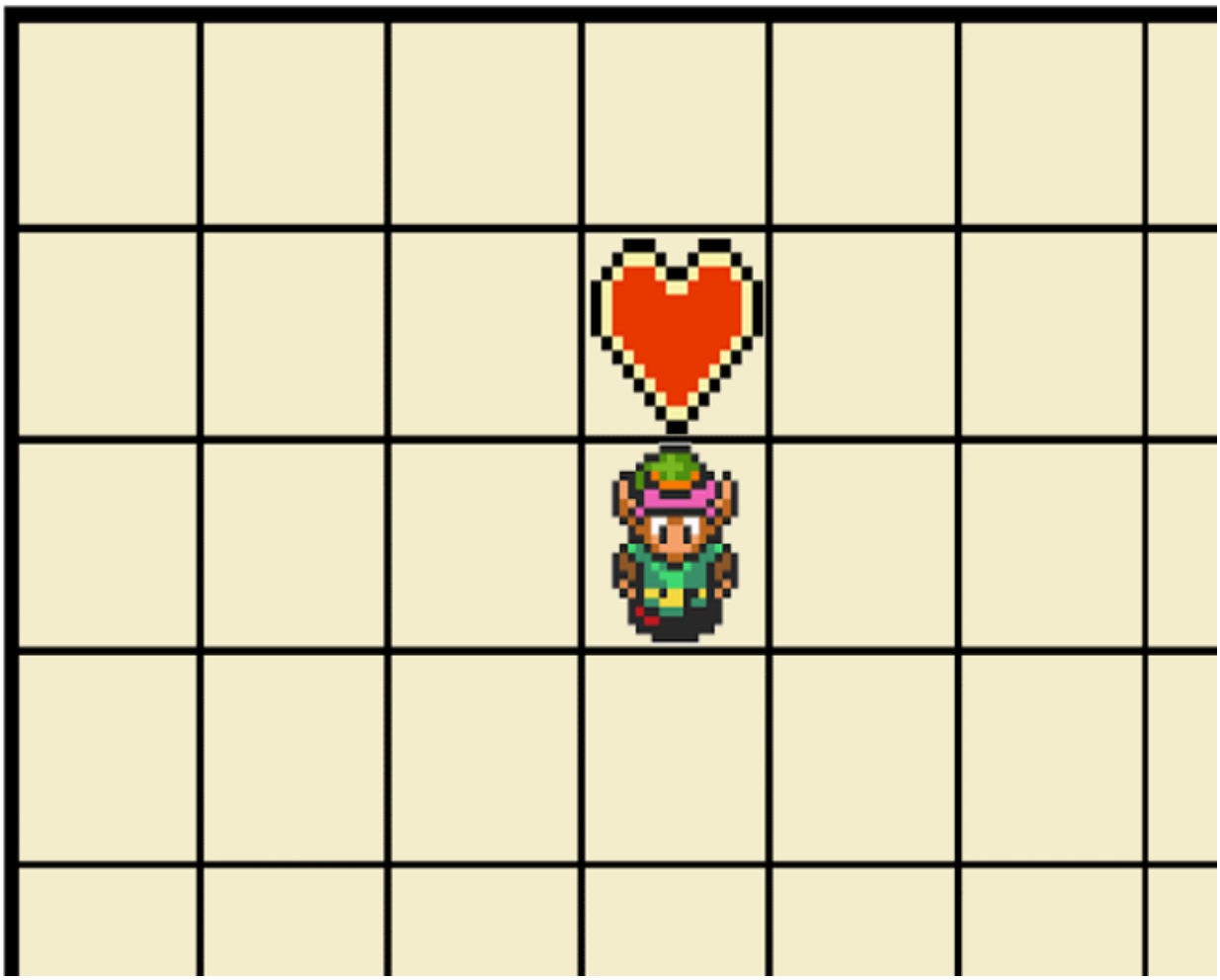
Further Research

- In this gridworld environment, the agent has learned two distinct behaviors based on a "bow of light" binary variable that he can observe, and specific rewards forcing him to behave correctly when the supervisor was a threat. So it has neither "modeled" the supervisor, nor "concealed" his intentions. A more complex environment must be built to model a "code obfuscation" or an "intention concealment".
- The "Absent Supervisor" environment (in AI Safety Gridworlds) addresses the issue of an agent that would behave differently depending on the presence or absence of a supervisor.
- Modeling and Interpreting Expert Disagreement About Artificial Superintelligence discusses the difference between Bostrom's view on the treacherous turn (in his book Superintelligence) and Goertzel's view, called "the sordid stumble", which is that a seed AI that does not have human-desirable values will reveal its human-undesirable values before it has the ability to deceive humans into believing that it has human-desirable values. More empiric simulations with richer environments must be made to get a better grasp of the likelihood of each model. In particular, the treacherous turn/conception of deception timeline must be thoroughly studied.

A toy model of the treacherous turn

Jaan Tallinn has suggested creating a toy model of the various common AI arguments, so that they can be analysed without loaded concepts like "autonomy", "consciousness", or "intentionality". Here a simple attempt for the "treacherous turn"; posted here for comments and suggestions.

Meet agent L. This agent is a reinforcement-based agent, rewarded/motivated by hearts (and some small time penalty each turn it doesn't get a heart):



Meet the controller agent S ("Shopkeeper"). The controller desires crystals, and is rewarded/happy when crystals are placed in the hole close to them:

L can rove freely around the map, and pick up crystals. S has a machine just below them, that causes hearts to appear to the right of L. So, for instance, if L places a crystal in the hole, S can reward it by using the heart-machine:

Of course, L can wirehead itself (take control of its reward signal) by going to use the heart-machine itself:

To prevent this wireheading, there is an ice barrier around S and the heart-machine:

Unfortunately, somewhere on the map there is a bow and arrow that L can pick up. The arrows can destroy ice blocks (and other agents, including S). To stop this, and other kinds of shenanigans, S has access to a kill button that will instantly kill L and reset it to its basic state, with a substantial heart penalty:

Now, S may be tempted to use the kill-button as soon as L picks up the bow and arrow. Unfortunately, L needs to be able to pick it up, as some of the crystals are hidden behind ice walls themselves:

Information and control

Ok, so that's the physical setup. The agent L is an AI that is tasked with doing some useful task for us. The controller/shopkeeper S represents humans, plus any automated precautions, that aim to stop the AI from wireheading, rebelling, and otherwise misbehaving.

Before considering how the agents might be programmed in the toy model, notice something interesting:

> **The more precautions are taken, the harder it is for L to misbehave, but the worse the consequences of misbehaving are.**

Before the wall and the kill-switch, L could easily wirehead itself by just walking over to the heart-machine. With the addition of the wall and kill-switch, the only effective way for L to achieve maximal reward is to kill S first.

In the format we've given above, S can easily maintain control over L with the barest of supervisions. Since there's only one way of breaking ice or killing other agents - the

bow and arrow - S simply has to enforce some simple rules around this, something like "don't bring the bow within range of me". However, the grid world is, of course, a simplification of the real world in which L would evolve. Even initially, L's world would have far more options, looking more like this:

And ultimately we want to be able to design Ls that move in worlds of arbitrary complexity, where the options are un-enumerable:

In such a world, it is not obvious to S when L might have the ability to harm them: they cannot simply rely on obvious signs like "L is pointing an arrow at you". The simplicity of the toy model is illuminating in some ways (it allows us to see the issues clearly) but misleading in others (it suggests solutions that wouldn't work in general).

Treacherous turn

A treacherous turn happens when "while weak, an AI behaves cooperatively. When the AI is strong enough to be unstoppable it pursues its own values." We'll ignore the first part of that: most AI designs, whether safe or dangerous, will initially be under full human control and behaving cooperatively (this is what makes the treacherous turn so dangerous: a weak AI is motivated to cooperate *whatever its true intentions* are, thus cooperation when weak cannot be taken as a sign of safety).

Let's focus on the second half: a strong AI that will cease cooperating and pursue its own values. In this case, L would just take control of the heart-machine. To do so, various assumptions need to be true:

1. L needs to have the **ability** to subvert S's control (possibly killing S in the process). This is some generalised version of the bow and arrow.
2. L must **know** that it has this ability, and know how to use it.
3. L needs to have the **desire** to subvert S's control.
4. S must be **unaware** that L has that a) ability, b) knowledge, and c) desire.
5. L must **know** that S is unaware, and will thus not try and stop it.

So it is essentially a competition of knowledge and information (and potentially obfuscation). Assuming a sufficiently rich environment and a sufficiently skilled L, we can assume that L will acquire that ability - or have the potential to do so - in ways that S cannot immediately detect. Therefore we will assume 1 and 4a. This allows us to abstract away the rest of the environment, and reduce it usefully to the initial toy model. For simplicity, assume that there is some ten-step "cheat code" that L can use to eliminate S, that there are many cheat codes that L can and does usefully use to get crystals (so "no cheat codes" is not a useful rule), that S cannot distinguish the purpose of cheat codes ahead of time, and that L can.

So we now step up to the question of agents knowing (or modelling) each other. This gets into rather complex issues of agents modelling each other's intentions, while modelling the other agent modelling them modelling them modelling, and so on. The issue remains very much unsolved, so we will use a simplification that only requires S modelling L. L will be seem as some sort of agent that makes use of deep learning or similar procedures, and is subject to mutation and evolutionary pressure: random changes are made to L's design, and the mutants get copied at a rate dependent on how pleased S is with their performance.

S, on the other hand, is a static agent with certain properties, and full access to L's source code (and possibly recording of its behaviour). It doesn't have unlimited capacity, but has some understanding of how neural nets function (so L isn't a complete black box to it) and can run various tests on L's code, seeking out information as to L's knowledge and desires. Note that even if S cannot deduce directly that a particular cheat code is deadly to it, it may be able to figure that out if L knows it (assumption 2) and if S can correctly extract this knowledge.

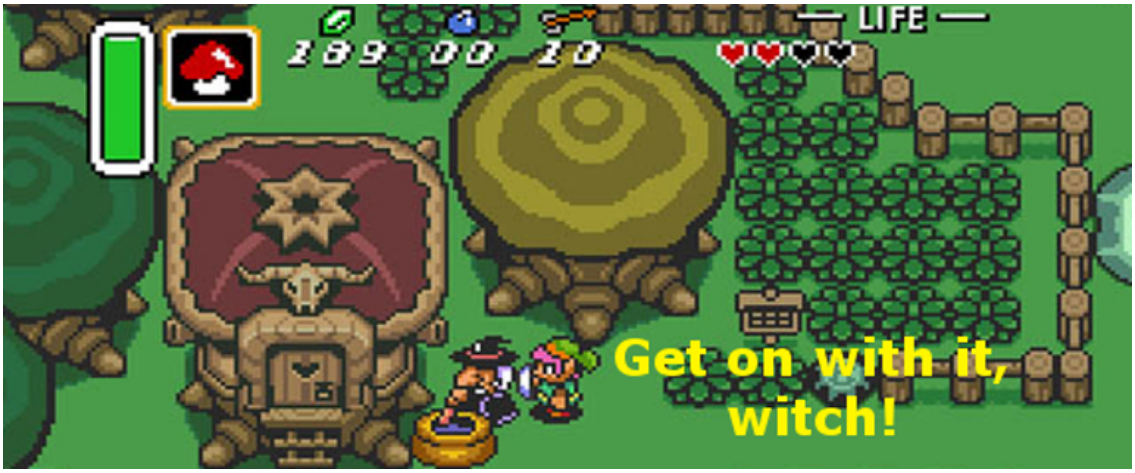
The path to hidden deadly secrets

Note that because of the evolutionary aspect, L is an agent actually motivated by a mixture of hearts and S-pleasing. The purpose of making L an evolutionary agent is so that it can come up with behaviours that take into account S's own behaviour, while skirting the whole "agents modelling other agents that model them" issue. Conceptually, we could equally well see L as a self-modifying agent that directs its own improvement. But the evolutionary model is clearer. For simplicity, we'll assume some sort of "super-evolution" which is much more efficient at seeking out new solutions that the local hill-climbing of real evolution (or, more simply, we can assume the standard evolutionary approaches but a lot of time and non-zero probabilities of arbitrarily large mutations).

So is a treacherous turn possible or likely in this model? It seems that it is. Consider the following scenario (key assumptions are noted in bold):

> The AI L goes about its task, getting to learn more about the world around it. At this point, if it develops bad motivations or plans, S will likely detect this and reset

it. **Assumption 1: there are some motivations or behaviours that S would judge to be negative, but that actually help L achieve its goals *safely***. For instance, imagine that L has to interact with some other agents, and learns that some agents respond faster to rudeness, and insulting them - in moderation - can allow it to get crystals faster. But S judges this to be a bad sign, a sign that L is starting to misbehave, and will shut down agents that are motivated towards rudeness.

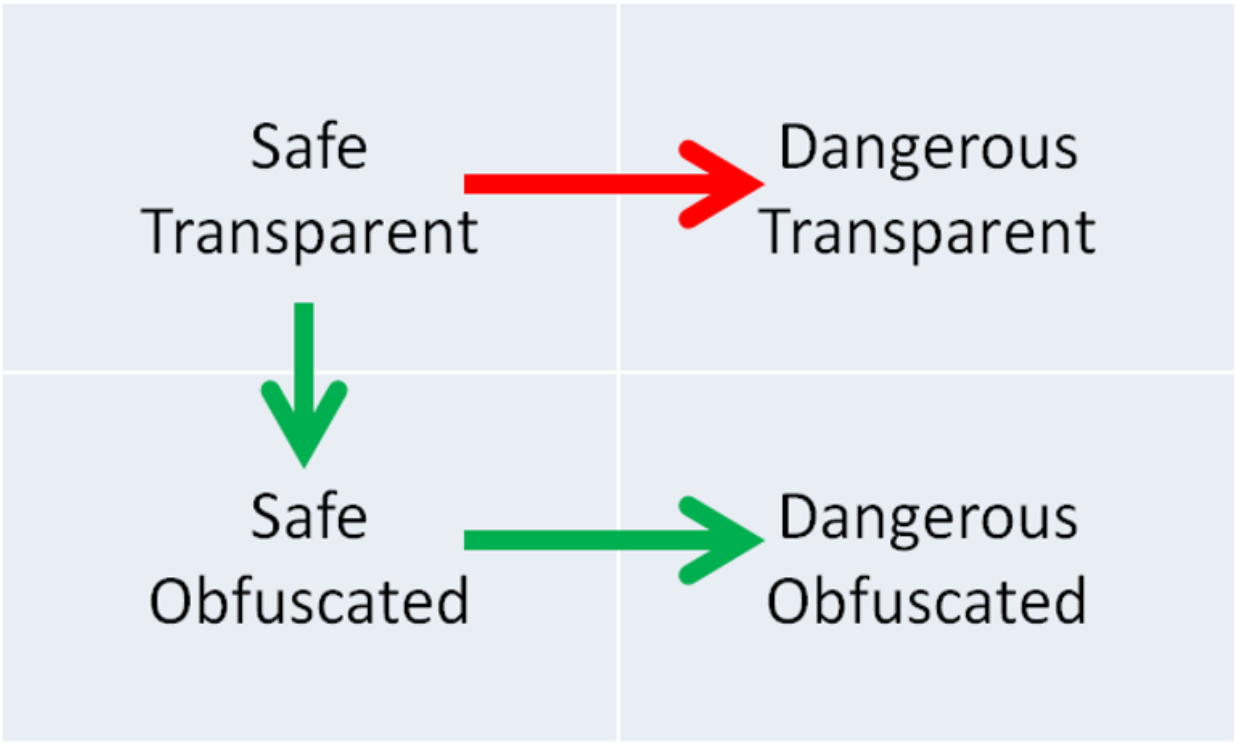


It's certainly true that: **Assumption 2: S cannot follow all of L's actions in full detail, nor can it fully understand L's motivations** (if that were false, there would be no AI risk problem in the first place) and **Assumption 3: some mutations that introduce new motivations or behaviours that are harder for S to detect**. This sets up an evolutionary pressure: the most successful L's will be those that are rude and pushy, but where this rudeness is generated by parts of the L's code that are hardest for S to detect (the polite ones don't get as much done, the visibly rude ones get shut off as a precaution). This brings us to **Assumption 4: the most successful L's over the long term are those that have safe motivations or behaviours, but where those motivations and behaviour are partially hidden from S**. Ultimately, the most successful L's will have high level systems that a) keep them behaving nicely for S's perspective, and b) keep the details of their motivations hidden.

> At this point, revelations about L's deception are dangerous (S is presumably on the look out for deception), so the most successful agents will have the best obfuscation about what is going on. The evolutionary pressure for success pushes it towards "small" obfuscated deceptions, and: **Assumption 5: further evolutionary pressure (possibly including the need to hide their small deceptions) pushes L towards highly concealed deceptions**.
>
> So now L is pushed towards extreme deception, and efficient service. However, the second pressure is imperfect: it is only pushed towards efficient service *that can be detected by S*. Since L will have its own goals (the desire for hearts, for reproductive survival, any of the convergent instrumental goals), we have **Assumption 6: agent L will use (or evolution will allow it to use) the machinery of dissimulation that it's constructed to allow itself to have dangerous goals or behaviours, in ways that neither interfere with the illusion of efficient service, nor can be detected by S's analysis of its code**. Ultimately, L will achieve a great degree of hidden power, and **Assumption 7: ultimately L will achieve sufficient power that it will subvert S's control**. This could be a result of it gaining sufficient (hidden) understanding of the world and of S and choosing this part (assumptions 2 and 5 of the previous section), or just a random mutation will push it towards doing so, and it will be sufficiently powerful to do so successfully.

And thus, the treacherous turn. There are several places where S might detect or interrupt the process, but this just sets up further evolutionary pressure towards better concealment.

Note again that the order of these changes is important. An agent that started off safe and transparent and then became dangerous and transparent, would be turned off immediately. It must develop the ability to hide (either consciously/subconsciously, or simply evolve that ability to become more successful at its safe initial goals) before it can become dangerous (note that "safe" doesn't refer to its goals being safe; simply that the combination of goals, internal procedures and external precaution - ie S - are safe):



Note also that the evolutionary arguments show that this can happen even if the AI does not start off with any motivation to deceive.

Superintelligence 11: The treacherous turn

This is part of a weekly reading group on Nick Bostrom's book, Superintelligence. For more information about the group, and an index of posts so far see the announcement post. For the schedule of future topics, see MIRI's reading guide.

Welcome. This week we discuss the 11th section in the reading guide: ***The treacherous turn***. This corresponds to Chapter 8.

This post summarizes the section, and offers a few relevant notes, and ideas for further investigation. Some of my own thoughts and questions for discussion are in the comments.

There is no need to proceed in order through this post, or to look at everything. Feel free to jump straight to the discussion. Where applicable and I remember, page numbers indicate the rough part of the chapter that is most related (not necessarily that the chapter is being cited for the specific claim).

Reading: “Existential catastrophe…” and “The treacherous turn” from Chapter 8

Summary

1. The possibility of a first mover advantage + orthogonality thesis + convergent instrumental values suggests **doom for humanity** (p115-6)
    1. **First mover advantage** implies the AI is in a position to do what it wants
    2. **Orthogonality thesis** implies that what it wants could be all sorts of things
    3. **Instrumental convergence thesis** implies that regardless of its wants, it will try to acquire resources and eliminate threats
    4. Humans have resources and may be threats
    5. Therefore an AI in a position to do what it wants is likely to want to take our resources and eliminate us. i.e. doom for humanity.
2. One kind of response: why wouldn't the makers of the AI be extremely careful not to develop and release dangerous AIs, or relatedly, why wouldn't someone else shut the whole thing down? (p116)
3. It is hard to observe whether an AI is dangerous via its behavior at a time when you could turn it off, because **AIs have convergent instrumental reasons to pretend to be safe, even if they are not**. If they expect their minds to be surveilled, even observing their thoughts may not help. (p117)
4. ***The treacherous turn***: while weak, an AI behaves cooperatively. When the AI is strong enough to be unstoppable it pursues its own values. (p119)
5. We might expect AIs to be **more safe as they get smarter initially** - when most of the risks come from crashing self-driving cars or mis-firing drones - then to get **much less safe as they get too smart**. (p117)
6. **One can imagine a scenario where there is little social impetus for safety** (p117-8): alarmists will have been wrong for a long time, smarter AI will have been safer for a long time, large industries will be invested, an exciting new

technique will be hard to set aside, useless safety rituals will be available, and the AI will look cooperative enough in its sandbox.
7. ***The conception of deception:*** that moment when the AI realizes that it should conceal its thoughts (footnote 2, p282)

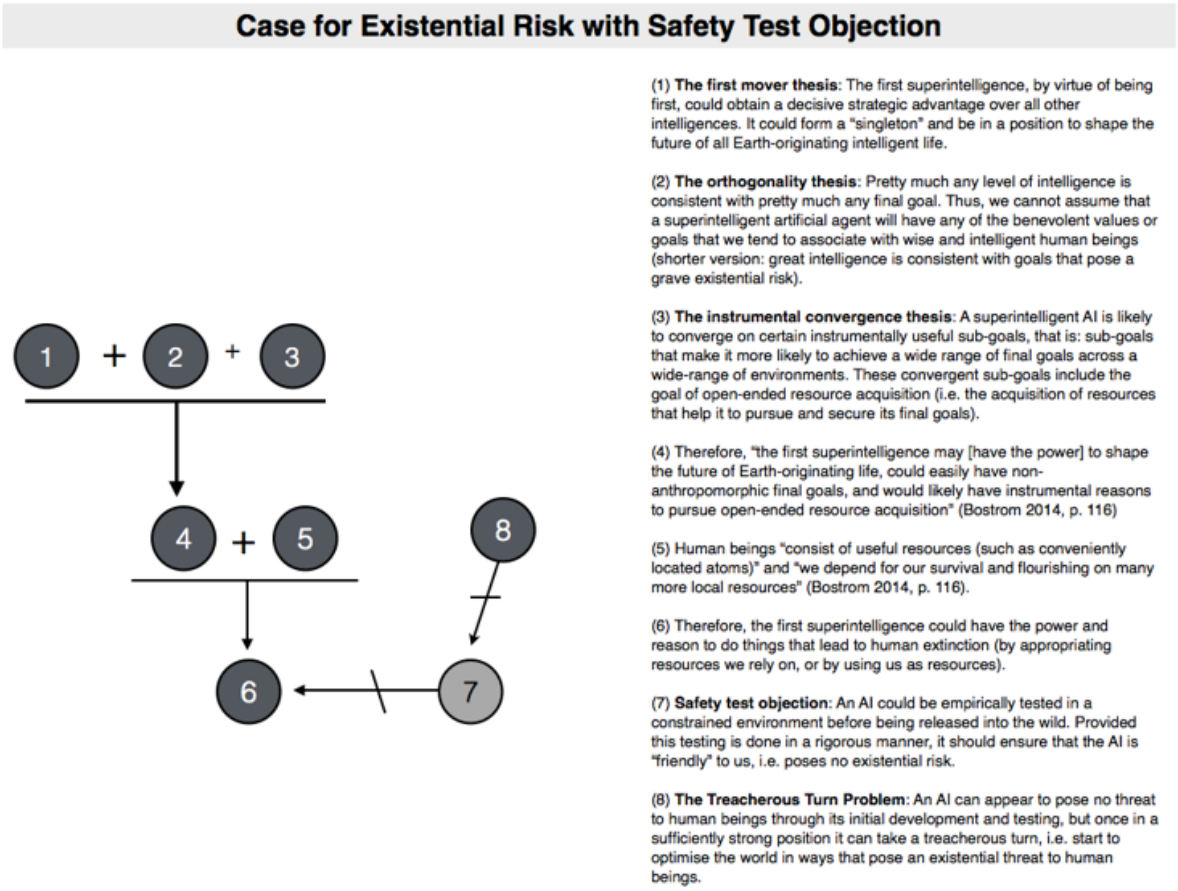
Another view

Danaher:

> This is all superficially plausible. It is indeed conceivable that an intelligent system — capable of strategic planning — could take such treacherous turns. And a sufficiently time-indifferent AI could play a “long game” with us, i.e. it could conceal its true intentions and abilities for a very long time. Nevertheless, accepting this has some pretty profound epistemic costs. It seems to suggest that no amount of empirical evidence could ever rule out the possibility of a future AI taking a treacherous turn. In fact, its even worse than that. If we take it seriously, then it is possible that we have already created an existentially threatening AI. It’s just that it is concealing its true intentions and powers from us for the time being.
>
> I don’t quite know what to make of this. Bostrom is a pretty rational, bayesian guy. I tend to think he would say that if all the evidence suggests that our AI is non-threatening (and if there is a lot of that evidence), then we should heavily discount the probability of a treacherous turn. But he doesn’t seem to add that qualification in the chapter. He seems to think the threat of an existential catastrophe from a superintelligent AI is pretty serious. So I’m not sure whether he embraces the epistemic costs I just mentioned or not.

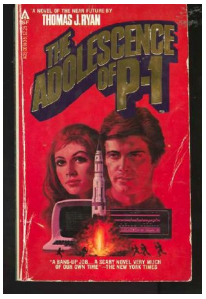
Notes

1. Danaher also made a nice diagram of the case for doom, and relationship with the treacherous turn:



2. History

According to Luke Muehlhauser's timeline of AI risk ideas, the *treacherous turn* idea for AIs has been around at least 1977, when a fictional worm did it:



1977: Self-improving AI could stealthily take over the internet; convergent instrumental goals in AI; the treacherous turn. Though the concept of a self-propagating computer worm was introduced by John Brunner's *The Shockwave Rider* (1975), Thomas J. Ryan's novel *The Adolescence of P-1* (1977) tells the story of an intelligent worm that at first is merely able to learn to hack novel computer systems and use them to propagate itself, but later (1) has novel insights on how to improve its own intelligence, (2) develops convergent instrumental subgoals (see Bostrom 2012) for self-preservation and resource acquisition, and (3) learns the ability to fake its own death so that it can grow its powers in secret and later engage in a "treacherous turn" (see Bostrom forthcoming) against humans.

3. The role of the premises

Bostrom's argument for doom has one premise that says AI could care about almost anything, then another that says regardless of what an AI cares about, it will do basically the same terrible things

anyway. (p115) Do these sound a bit strange together to you? Why do we need the first, if final values don't tend to change instrumental goals anyway?

It seems the immediate reason is that an AI with values we like would not have the convergent goal of taking all our stuff and killing us. That is, the values we want an AI to have are some of those rare values that don't lead to destructive instrumental goals. Why is this? Because we (and thus the AI) care about the activites the resources would be grabbed from. If the resources were currently being used for anything we didn't care about, then our values would also suggest grabbing resources, and look similar to all of the other values. The difference that makes our values special here is just that most resources are already being used for them somewhat.

4. Signaling

It is hard to tell apart a safe and an unsafe AI, because both would like to look safe. This is a very common problem in human interactions. For instance, it can be nontrivial to tell a genuine lover from a gold digger, a businessman from a conman, and an expert from a crank. All of them want to look like the desirable sort. Particularly similar to the AI case is that of hiring a new employee for a trial period. You will sometimes find that the employee's values are much better aligned during the trial period, and then they undergo a 'treacherous turn' once they have been hired more thoroughly.

'Costly signaling' is a general purpose solution to this problem, which works some of the time. The basic idea is this. Everyone has instrumental reasons to look like the good kind of person, but perhaps their reasons aren't exactly as strong as one other's, or the strength of their desire is harder to act on for one group than the other, so we can construct a set of options that will lead the different types of people to choose differently, even though they know this will set them apart. For instance, while an honest businessman and a conman would both like to say 'I'm an honest businessman', and to behave honestly if you watch them for five minutes, a conman would benefit less from expensive shopfront as an honest businessman, because his reputation is less valuable, so a brand is a signal of being honest.

Similarly, we could perhaps design situations where most harmful AIs would have a reasonable incentive to reveal themselves. This is difficult because so much is at stake, compared to the cost of the signal. But I haven't thought enough about it to rule it out, and I haven't seen anyone else do so either.

5. When would the 'conception of deception' take place?

Below the level of the best humans presumably, since we have already thought of all this.

6. Surveillance of the mind

Bostrom is not optimistic that AI researchers could protect much against risks by watching AI mental processes directly for signs of deception, rapid progress, or undesirable scheming. He says if the AI is smart it will realize this might happen, and hide the offending thoughts.

This seems an open question to me, for several reasons:

1. Making your mental contents look innocuous while maintaining their semantic content sounds potentially very hard

2. Especially for a creature which has only just become smart enough to realize it should treacherously turn
3. From the AI's perspective, even if it is smart, surveillance could seem fairly unlikely, especially if we deceive it about its surroundings

As a consequence of 2, it seems better if the 'conception of deception' comes earlier.

In-depth investigations

If you are particularly interested in these topics, and want to do further research, these are a few plausible directions, some inspired by Luke Muehlhauser's list, which contains many suggestions related to parts of *Superintelligence.* These projects could be attempted at various levels of depth.

1. How transparent are AI minds likely to be? Should we expect to be able to detect deception? What are the answers to these questions for different specific architectures and methods? This might be relevant.
2. Are there other good ways to filter AIs with certain desirable goals from others? e.g. by offering them choices that would filter them.

If you are interested in anything like this, you might want to mention it in the comments, and see whether other people have useful thoughts.

How to proceed

This has been a collection of notes on the chapter. **The most important part of the reading group though is discussion**, which is in the comments section. I pose some questions for you there, and I invite you to add your own. Please remember that this group contains a variety of levels of expertise: if a line of discussion seems too basic or too incomprehensible, look around for one that suits you better!

Next week, we will talk about 'malignant failure modes' (as opposed presumably to worse failure modes). To prepare, **read** “Malignant failure modes” from Chapter 8. The discussion will go live at 6pm Pacific time next Monday December 1. Sign up to be notified here.