Superintelligence and Control, from the beginning

- *Turing (1951)* about the First formulation of the control problem: "it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. At some stage therefore we should have to expect the **machines to take control**."
- *Good (1965)* started to define singularity with improvement: "let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even *better machines (recursive improvement)*; there would then unquestionably be an intelligence explosion, and the intelligence of man would be left far behind. Thus, the first **ultra-intelligent** machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

In these quotations, there are important arguments: the idea of an explosion to some ***recursive improvement*** and the fact that a ***singularity*** can be reached. There is also mentioned the idea of alignment to avoid the control problem.

[The idea of recursive improvements is very similar to the idea of Open AI as a long strategy: on the Open AI site is contained a paper under the alignment strategy. The main idea of the strategy is that we are going to use an unaligned AI against the alignment problem]

- We will build AIs that are much more intelligent than humans (i.e.: **super-intelligent**).
- Those AIs will be autonomous agents with pursue large-scale goals.
- Those goals will be **misaligned** with ours; that is, they will aim towards outcomes that aren't desirable by our standards and trade-off against our goals.
- The development of such AIs would lead to them gaining control of humanity's future (as an existential risk).

Questions:

1. What is super-intelligent AI defined and is it reasonable that we get there?
2. Even if we get there is it reasonable to assume that those agents pursue large-scale goals?
3. Even if we grant the previous promises, is it reasonable to expect that such an agent will have aims that are misaligned with ours?
4. Even if we grant the previous promises, how can we prevent it?

Narrow vs General AI

- ***Narrow***: it is *task-based*, agents that understand how to do well at many tasks because they have been specifically optimized for each task.
- ***General***: it is *generalized-based*, agents which can understand new tasks with little or no task-specific training, by generalizing from previous experience, and displaying good performances on previously unseen tasks.

It has to be considered as a part of a spectrum rather than a binary classification, particularly because the way we choose how to divide up tasks can be quite arbitrary.

Narrow (task-based approach)

Evolution trained us to perform some cognitive skills, certain sensory and processing skills, and social skills, and then we harnessed (sfruttare) skilling them up to do very well and very complex that are not naturally in our evolutionary training.

For instance, humans harnessed and used ***electricity***: while electricity is a powerful and general technology, we still need to design specific ways to apply it to each task. Or also ***computers***: these are powerful and flexible tools but even though they can process arbitrarily many different inputs, detailed instructions for how to do that processing needs to be individually written to build each piece of software.

Our current reinforcement learning algorithms, although powerful, produce agents that are only able to perform well on specific tasks at which they have a lot of experience.

Drexler argues that our current **task-based approach** will scale up to allow superhuman performance on a specific range of complex tasks.

General (generalization-based approach)

As a ***species***, we were trained by evolution to have cognitive skills including rapid learning capabilities, sensory and motor processing, and social skills. As ***individuals***, we were also trained during our childhoods to fine-tune those skills, to understand spoken and written language, and to possess detailed knowledge about modern society. However, the key point is that almost of all this evolutionary and childhood learning occurred tasks on the economically useful ones we perform as adults.

The **skill of abstraction** allows us to extract common structure from different situations, which allow us to understand them much more efficiently than by learning about them one by one (it is called **generalization-based approach** because AI can generalize from previous experience). Then our communication skills and theories of mind allow us to share our ideas. This is why humans can make great progress on the scale of years, not just via evolutionary adaptation over many lifetimes.

Large language models like *GPT-2* and *GPT-3* are an example of generalization-based approaches.

Comparison between narrow and general

The *task-based approach* (*narrow intelligence*) is useful when we get and use lots of data. It is very useful in demanding professions like medicine, law and mathematics.

However, some jobs depend on the ability to analyze and act on a wide range of information that it will be very difficult to train directly for high performance on them. For instance, the role of CEO in a large company: this requires a particular vision and feeling of interpreting what you see and will happen; it is not based only on specific data. It is not a narrow intelligence task, but it *requires a general intelligence*.

Ngo (2020) stated: "eventually (one day) we will be able to create AIs that can ***generalize*** well enough to produce human-level performance on a wide range of tasks, including abstract low-data tasks like running a company. We call these systems artificial general intelligence or **AGIs**".

Is it reasonable that we will be able to get there?

Bostrom (2014) defines a ***superintelligence*** as *"any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest"* (doing better than all of humanity could if we coordinated globally).

It's difficult to deny that in principle it's possible to build individual generalisation-based AGIs which are superintelligent, since human brains are ***constrained*** by many factors which will be much less limiting for AIs, for instance:

- **Human brains are constrained** by many factors which will be much less limiting for AIs.
- The difference between the speeds of **neurons** and **transistors**: the latter pass signals about four million times more quickly.
- A neural network can be several orders of **magnitude** larger than human brain.
- Evolution has not had much time to select specifically for the skills that are most useful in our modern environment, such as linguistic competence and mathematical reasoning.

Transition from human-level AGI to superintelligence

By default, we should expect that the path that led to the superintelligence will be driven by the standard factors which influence progress in AI: *more compute*, *better algorithms* and *better training data*.

However, there are three important factors that contribute to increase AI intelligence:

1) - **Replication**: AIs are much less constrained than humans, so it is very easy to create a duplicate of an AI which has all the same skills and knowledge as the original. The cost of compute for doing so is likely to be many times smaller than the original cost of training an AGI.
Duplicating such an AGI could give rise to a superintelligence composed not of a single AGI, but rather a large group of them. Furthermore, these collective AGIs can carry out more **complex tasks** than the original, thanks to the composition of difficult tasks in subtasks.
We expect superintelligence as these collective AGIs (however, there could be coordination problems between its members).

2) - **Cultural Learning:** We should expect AGIs to be able to acquire knowledge from each other and then share their own discoveries in turn, allowing a collective AGI to solve harder problems than any individual AGI within it could. The development of this ability in humans is what allowed the *dramatic rise of civilisation* over the last ten thousand years.

3) - **Recursive Improvement:** AGIs will be able to improve the training processes used to develop their successors, which then improve the training processes used to develop their successors and so on.

Due to the ease of duplicating AIs, then there is no meaningful distinction between an AI improving itself versus creating a successor that shares many of its properties.

Modern AIs are more accurately characterized as models which could be retrained, rather than software that could be rewritten. Almost all of the work of making a neural network intelligent is done by optimiser via extensive training more accurately to think about self-modification as the process of an AGI modifying its high-level architecture or training regime, then putting itself through significantly more training.

Goals and Agency

How will they decide what to do actually? For instance, will the individuals within a collective AGI even want to cooperate with each other to pursue larger goals? AIs will gain too much power over humans, and then use that power in ways we don't endorse. Why might they end up with that power? There are *three possibilities*

1. AIs pursue power for the sake of achieving other goals, i.e. **power is an instrumental** goal for them.
2. AIs pursue power for its own sake, i.e. **power is a final goal** for them.
3. AIs gain power without aiming towards it, i.e. because **humans gave it to them**.

1. Power as an instrumental goal

The key idea behind the first possibility is *Bostrom (2012)'s* **instrumental convergence thesis**, which states that there are some instrumental goals whose attainment (raggiungimento) would increase the chances of an agent's final goals being realised for a wide range of final goals. There are some instrumental goals likely to be pursued by almost any intelligent agent, because there are some objectives that are useful intermediaries to the achievement of almost any final goal. Examples of such instrumental goals (which are useful for executing large-scale plans) are:

- *Self-preservation*
- *Resource acquisition*
- *Technological development*
- *Self-improvement*

However, the link from instrumental goals to dangerous influence-seeking is only applicable to agents which have final goals large-scale enough to benefit from these instrumental goals. It is not yet clear that AGIs will be this type of agent or have this type of goals (gain power as an instrumental goal).

Our conquest of the world didn't require any humans to strategize over the timeframe of centuries, but merely for many individuals to expand their personal influence in a **relatively limited way**.

Furthermore, we should take seriously the possibility that superintelligent AGIs might be even less focused than humans are on achieving large-scale goals.

But what does it mean to be an agent?

We have to distinguish between the goals which an agent has been selected to do well at (its **design objectives**) and the goals which an agent itself wants to achieve (**agent's goals**).

What does it mean for an agent to have a *goal for its own*? There are three different approaches that try to answer this question:

- By Morgenstern and Von Neumann means *expected utility maximisation*
- By Dennett's intentional stance (posizione): taking the *intentional stance towards the systems* can be useful for making predictions about them (this only works given the prior knowledge about what goals they are most likely to have)
- By Hubinger means *mesa-optimisation*

Alignment

So, the argument is composed by four points:

1. the first about ***clarifying when we will get to super intelligent agents***
2. ***how these superintelligent agents might pursue goals*** on their own (instrumentally or not)
3. ***whether these goals are misaligned***.

There are reasons to worry that AIs will develop undesirable final goals which leads to conflict with humans. But what does alignment (with human values) mean? There are two ways to describe the *meaning* of **alignment**:

- **Minimalist** (*narrow definition*) approaches focus on avoiding catastrophic outcomes. The best example is *Christiano's* (2008) concept of intent alignment: when I say that an AI is aligned with an operator H, I mean that A is trying to do what H wants it to do.

- **Maximalist** (*ambitious definition*) approaches attempt to make AIs to adopt a specific overarching set of values (like moral theories, global democratic consensus or meta-level procedure for deciding between *moral theories*). In this sense, we can have maximally or totally aligned agents.

The **maximalist definition** blends together so many different levels (*moral, social, political*…) and this requires a level of technological development that actual we don't have.

The most studied and most used is the **minimalist (narrow) definition** that tries to avoid catastrophic outcomes. This approach states that *"an AI is said to be misaligned with a human if the human would want the AI not to do what AI is trying to do (if the human is aware of AI's intentions)".*

Thus, we focus on the **minimalist definition** and about it we can make some considerations:

- This definition implies that AIs could potentially be neither aligned nor misaligned with an operator (for instance, if they only do things which the operator does not care about).
- By using the word "*trying*", they focus on the ***AI's intentions****,* not on the actual outcomes achieved

EXAMPLE WITH A TOY: THE TYPE OF TOY YOU WANT TO PLAY IS SEPARATE FROM THE WAY YOU PLAY WITH IT.

- MIGTH WANT A TOY CAR BUT CAN BE USED IN DIFFERENT WAYS (RACING ON THE FLOOR OR TO PRETEND A CAR WASH)

CHESS EXAMPLE: PROGRAM THAT PLAY CHESS: THE GOAL IS WINNING

TO THIK THE MOVES ONE PROGRAM COULD MAKE A LOT OF COMPLEX COMPUTATION, ANOTHER ONE COULD TRY TO FIND PATTERS. THE WAY THE PROGRAM THINKS FOR REACHING ITS ULTIMATE GOAL DOES NOT CHANGE ITS ULTIMATE GOAL (WIN GAME)

- So, the agents don't misbehave just because they mis-understand what we want or interpret our instructions overly literally (it is called *perverse instantiation*) but AGIs will understand what we want and ***just not care***, because the motivations they acquired during training weren't those we intended them to have.

!! The idea that AIs won't automatically gain the right motivations by virtue of being more intelligent is an implication of *Bostrom's (2012)* **orthogonality thesis**: any level of intelligence could in principle be combined with any final goals. Consider any high-functioning psychopaths, who understand that other people are motivated by morality and can use that fact to predict their actions and manipulate them, but nevertheless are not motivated by morality themselves. (Here the idea is not that when something becomes more intelligent, then it becomes less stupid and it is less evil, on the contrary the idea is that intelligence could be combined with any final goals, **also evil aims**).

Outer and Inner misalignment: Standard Picture

We train machine learning systems to perform desired behaviour by optimising them with respect to some objective function, for instance a *reward function* in reinforcement learning. There could be considered two types of misalignments:

a) - **Outer misalignments**: concern is that we won't be able to implement an objective function which describes the behaviour we actually want the system to perform, without also rewarding misbehaviour.

b) - **Inner misalignment**: our agents might develop goals which differ from the ones specified by that objective function. This is likely to occur when the training environment contains subgoals which are consistently useful for scoring highly on the given reward function, such as gathering resources and information, or gaining power.

a) Outer misalignment

Problem: we won't be able to implement an objective function which describes the behaviour we actually want to the system to perform, without also rewarding misbehaviour. But why is it **difficult** to specify objective functions?

- Difficulty of explicitly programming objective functions **which express all our desires** about AGI behaviour
- There is no simple metric that we'd like our agents to maximise – rather, desirable AGI behaviour is best formulated in terms of concepts like obedience, consent, helpfulness, morality and cooperation (which we **can't define precisely** in realistic environments)
- *Goodhart's law* suggests that some undesirable behaviour will score very well according to **proxies** (deleghe) we might define for those concepts, and therefore be reinforces in AIs trained on them. (E.g.: If there is a test, and you know the test in advance, you are going to train for the test and no for the knowledge in advance).

How to address these problems?

WHEN WE TEACH AIs HOW TO DO COMPLEX TASKS (DRIVING A CAR, DIAGNOSING A MEDICAL CONDITION) WE WANT TO MAKE SURE WE ARE DOING IT CORRECTLY. BUT IT WOULD BE TOO MUCH WORK FOR HUMANS TO CHECK ALL THE DECISIONS THE AI TAKES

One idea is by incorporating ***human feedback*** into the objectives function used to evaluate AI behaviour during training. However, this approach could face some challenges:

- It would be prohibitively expensive for humans to provide feedback on all data required to train AIs on complex tasks. This is known as the **scalable oversight problem**: reward modelling is the primary approach to addressing it.
- For long term tasks, we might need to give feedback before we have had the chance to see all the consequences of an agent's actions. Yet even in domains as simple as GO, it is often very difficult to determine how good a given move is without seeing the game play out. And in larger domains, there may be **too many complex consequences** for any single individual to evaluate. The main approach to addressing this issue is by using *multiple AIs* to recursively decompose the problem of evaluation.
- Humans can be manipulated into **interpreting behaviour more positively** than they otherwise would, for example by giving them misleading data.

b) Inner misalignment

Our agents might develop goals which differ from the ones specified by that reward function:

- The training environment contains **subgoals** which are consistently useful for scoring highly on the given objective function, such as gathering resources and information, or gaining power.
- If agents reliably gain higher reward after achieving such subgoals, then the optimiser might select for agents which care about these subgoals for their **own sake**

Humans, when we are trained by evolution to increase our genetic fitness. In our ancestral environment, subgoals like *love, happiness and social status* were useful for achieving higher inclusive genetic fitness, and so **we evolved to care about them**. But now we are powerful enough to reshape the natural world according to our ***desires*** and so there are significant differences between the behaviour which would maximize genetic fitness (e.g., frequent sperm or egg donation) and the behaviour which we display in pursuit of the motivations we actually evolved.

The idea expressed above regards how some instrumental subgoals could pursuit for own sakes or could **distract an agent from the real goal**.

[Suppose we reward an agent every time it correctly follows a human instruction; so that the cognition which leads to this behaviour is reinforced by its optimiser. Intuitively, we'd hope that the agent comes to have the goal of obedience to humans. But it is also conceivable that the agent's obedient behaviour is driven by the goal 'don't get shut down', if the agent understands that disobedience will get it shut down – in which case the optimiser might actually reinforce the goal of survival every time it leads to a completed instruction. So, two agents, each motivated by one of these goals, might behave very similarly until they are in a position to be disobedient without being shut down. One important factor is whether there are subgoals which reliably lead to higher reward during training. Another is how easy and beneficial it is for the optimiser to make the agent motivated by those subgoals, versus motivated by the objective function it is being trained on.]

So **how can we ensure inner alignment** of AGIs with human intentions?

One approach involves ***adding training examples*** where the behaviour of agents motivated by misaligned goals diverges from that of aligned agents. Designing and creating this sort of **adversarial training data** is currently much more difficult than mass-producing data for three reasons:

- Firstly, we simply won't know which **undesirable motivations** our agents are developing, and therefore which ones to focus on penalising. Interpretability techniques could help with this approach, but seem very difficult to create
- Secondly, it is very hard to add these training examples because the misaligned motivations which agents are most likely to acquire are those which are **most useful.** It is difficult to create such an environment where access to more information lead to lower rewards.
- Thirdly, we are most concerned about agents which have **large-scale misaligned goals**. Yet large-scale scenarios are again the most difficult to set up during training either in simulation or in the real-world.

Wireheading problem (given some implicit goal G, an agent wireheads if, instead of moving towards G, it manipulates some narrow measurement channel that is intended to measure G but will fail to do so after the agent's manipulation) regards using of *reward function*: the idea of reward function is approx to really what you want to reach.

For instance, you have children, and you want to sure that they get good grades. If they get good grades at the end of the year, you will buy them a motorbike. So, in order to get motorbike, they should work hard during the year, but what they want to maximize is having good grates. The main problem with this type of reasoning is that the **children try to get the scooter not good grades**. If the children are really smart, one way to get good grades is not studying more but there are other ways (children avoid the problem of studying more inventing new ways to get good grades).

In a sense there is a certain mistake between the **message** and the **channel**. Wireheading is the main problem regarding the reward function.

It is also possible that the agent learns to care about the state of the channel itself. Pain in animals is one example of this. The ***message*** is that damage is being caused; the ***code*** is that more pain implies more damage (as well as other subtitles of type and intensity), and the ***channel*** is the neurons that carry those signals to our brain about something has been damaged.

The same regards humans: at the end of the day, we care about as much the channel as we care about the message. Why? Because we want to avoid pain even if there are no damage in our body. This is shown by an experiment using *electroshock*: even if there are not damages, we try to prevent these signals because they are painful (we care about the **channel rather than the message**).

Similarly, an agent which was trained via a reward signal may desire to (?) continue receiving those signals even when they no longer carry the same message.

Control

Topic of control (4) represents the fourth point in the argument. The four points of the arguments are:

1. It is about ***clarifying when we will get to super intelligent agents***
2. It is about ***how these superintelligent agents might pursue goals*** on their own (instrumentally or not)
3. It is about ***if these goals are misaligned with ours***
4. It is about ***control***

If we fail concrete super intelligent agents whose large-scale goals are aligned with ours, still to be established whether they succeed in **taking control over humanity future**: that is the problem of control.

There are two kinds of disasters scenarios on the control problem:

1. **Conservative**: involves artificial intelligent agents slightly gaining influence within our current, political, economic system by taking control of companies or institutions.
   The idea is that we just reach a certain point where these AGIs are no longer incentivized to f*ollow human laws*. How could this happen? Humans might lose influence because they have less influence on certain strategic tasks (wars, climate change, political issues…) but no single AGI will actually control the whole world.
   That is why it is called the conservative scenarios: nothing structural really changed, it is analogous to how large corporation and institution accumulate power (even when most humans disapprove of their goals).

2. **Disrupting:** singularity. A single AGI will be able to gain enough power via such breakthroughs (scoperte) that they can seize control of the world. However, this purpose represents more a fantastic/imaginary scenario because *potential future technologies*, which would provide a decisive strategic advantage if possessed only by a single actor, don't yet exist.

What are the factors that will have an influence on us remaining in control? There are 4 factors:

1) - **Speed of development**: if the AI development proceed very quickly, then then our ability to react appropriately will be much lower. We should be interested in how long it will take for AGIs to proceed from human-level intelligence to superintelligence ***(take-off period).*** But how long has to be the so called '*take-off*' period? Consistently on what happened on the history systems (AlphaGO, AlphaStar) this period will be **very short**: after a long development period, each of them was able to improve rapidly from top amateur level to superhuman performance. An example is human evolution: it only took us a few million years to become much more intelligent than chimpanzees. However, it is unclear when this speed-up period arrives. Furthermore, there are also some criticisms: (1) there would be *gradual improvement* instead of a rapid speed-up (once reached the maximum development), (2) available of *computer power*

2) - **Transparency:** Increase transparency because if what is going on these AGIs we can actually access, we can predict how the system could be more transparency and more confident. But how to do that?
We can build *interpretability tools* that allow us to analyse the internal functioning of an existing system.
Or we can also create *training incentives* towards transparency. For instance, we might reward an agent for explaining its thought processes, or for behaving in predictable ways.
Another approach is to create and design algorithms that are intrinsic *more interpretable*.

3) - **Constrained deployment strategies**: A misaligned superintelligence with internet access will be able to create thousands of duplicates of itself, which we will have no control over, by buying (or

hacking) the necessary hardware. We can imagine trying to avoid this scenario by deploying AGIs in more *constrained ways* - for example by running them on secure hardware and only allowing them to take certain pre-approved actions.

4) - **Human political and economic combination:** Lastly, until it is not regulated and it is possible, If there is economic advantage we are going to increase economic and political combination. The idea about try to remain under control regards also trying to build economic consensus on how to deal: governments, companies, non-profits will vary in their responsiveness to safety concerns, cooperativeness, and ability to implement constrained deployment strategies. And the more of them are involved, the harder coordination between them will be.