How can AI and robots be combined so that they complement and contribute to our society, instead of posing a threat?

Towards safe, beneficial, human-enhancing AI

Silvia Santano, 2020

A commonly-accepted textbook definition of AI from “Artificial Intelligence: A Modern Approach”, by Stuart Russell and Peter Norvig, defines it as “designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment”[1] i.e., intelligent agents that act rationally to achieve the best possible outcome. A robot, on the other hand, is defined by the Oxford dictionary as a “machine—especially one programmable by a computer— capable of carrying out a complex series of actions automatically”. Artificial intelligence and robotics continuously motivate significant discussions about how they should be used, how they can be controlled and what risks they present. They have huge potential impact in numerous areas, such as healthcare, transportation, education and finance, among others. Thus, the immense socio-economic impact of robots and Artificial Intelligence on the development of humanity in the near future is ineludible. Unfortunately, its use also poses threats and challenges the humankind needs to overcome in order to be able to take advantage of them in a safe way.

The artificial intelligence in use today is properly known as narrow or weak AI and is designed to perform a specific problem solving task. Some of the tasks where it is already being implemented today include language translation, medical image analysis, recommender systems, face recognition, anomaly detection, virtual assistants or autonomous driving, to name a few. However, the long-term goal of many researchers is to create general AI (AGI). The hypothetical AGI would learn and outperform humans at nearly every cognitive task. The theoretical possibility of AI posing an existential risk generates a wide range of reactions both within the scientific community and in the public at large. There is a concept known as *singularity*, defined as a hypothetical point in time at which technological growth becomes out of control, resulting in unforeseeable und irreversible changes to human civilization. The first use of the concept in this context is believed to be from John von Neumann. Subsequent remarkable contributors have echoed this viewpoint. I. J. Good's "*intelligence explosion*"[2] (a possible outcome of humanity building AGI) model predicts that a future superintelligence will trigger a singularity. Stephen Hawking, Elon Musk, Steve Wozniak, Bill Gates, and many other popular figures in science and technology

[1] "Artificial Intelligence: A Modern Approach." Accessed October 23, 2020. http://aima.cs.berkeley.edu/.
[2] "Intelligence Explosion FAQ - Machine Intelligence Research ...." Accessed October 23, 2020. https://intelligence.org/ie-faq/.

have recently expressed concern[3] in the media and via open letters about the risks posed by AI, which they believe could possibly result in human extinction, and have been joined by many AI researchers. This view is, however, not shared by every member of the scientific community, as others assert that computers or machines will not achieve human-level intelligence and also among the ones asserting it will there are differences regarding when. Four polls of AI researchers, conducted in 2012 and 2013 by Nick Bostrom and Vincent C. Müller, suggested a median probability estimate of 50% that AGI would be developed by 2040–2050.[4]

When talking about superintelligent agents, the following question arises: can and should we be in control of something significantly smarter than us? Given the current pace of change and innovation, the question is how to leverage the benefits of artificial intelligence and robotics so that they contribute to our society, instead of posing a threat.

Let's explore possible approaches to some of the most imminent challenges with regards to AI broadly understood as any kind of artificial computational system that shows intelligent behavior.

Challenge 1: Maturity for Social Acceptance

A first barrier encountered before deploying AI systems in our society is the lack of the necessary acceptance by the society itself. Although remarkable differences can be found among individuals, partly affected by their education, affinity to and knowledge about technology, as well as background and previous exposure to it, as surveys show, there is a certain mistrustful perception of AI. One possible explanation for this is that subgroups that are more vulnerable to workplace automation express less enthusiasm for developing AI.
Indeed, one of the most extended, and perfectly understandable, fears is that related to jobs lost to automation.
Another fear, perhaps perceived as further away from the present but yet possible, is that of a world another step ahead, taken over by superintelligent robots (AGI) that make us humans sort of their slaves, in a similar fashion to how currently humans dominate animals. Another factor possibly having played a role in this perception is decades of science fiction depicting not so favorable consequences for humanity in imaginary scenarios.

---

[3] "Stephen Hawking warns artificial intelligence could end ...." Accessed October 23, 2020. https://www.bbc.com/news/av/embed/p02d9ysq/30290540.
[4] "Future Progress in Artificial Intelligence: A Survey of Expert ...." Accessed October 23, 2020. https://nickbostrom.com/papers/survey.pdf.

As Virginia Dignum points out in her “Ethics in artificial intelligence: introduction to the special issue”[5] article, these systems must be introduced in ways that build trust and understanding, and respect human and civil rights. Artificial Intelligence should be conceived as a complement to humans, not a substitute. The goal should be a society where people feel empowered, and not threatened by AI.
While it is true that already today AI systems might outperform us at concrete tasks, this doesn’t mean they need to replace humans. Instead, they should enhance humanity while still allowing them to flourish and without causing them harm. These topics and how to achieve them will be explored more in detail in the next sections.
In the following, I will outline some of the most prominent factors that might contribute to the current mistrust and ideas on how to approach them in their corresponding contexts. These are privacy, surveillance, explainability, empathy, machine ethics, responsibility and bias.

Privacy is definitely a major concern in artificial technologies, eliciting general discussion, mainly around the use of personal data. In a world where interaction with algorithms and robots is inevitable, the necessary privacy for citizens to feel safe ought to be guaranteed. Data collection, surveillance and even the analysis of human behavior of a specific individual to be used against them when taking decisions affecting their lives is a very delicate issue. There’s a difference between the advertisement industry, which avails itself of AI for targeted persuasion of potential consumers to purchase specific products or influence actions based on their behavior and apparent preferences, a practice which is already in a grey zone, and the abuse of this information with direct consequences for the person. To give an example, let’s imagine a machine is in charge of deciding whether a person is eligible for a loan or an insurance based on its personal preferences and camera footage of their whereabouts, where the person can be seen drinking and smoking. An argument can be made that all such practices represent an assault on personal privacy. While this would allow companies to make more informed decisions, this advantage should not be applied at all costs, especially not when it involves intruding in the private sphere of a human being. Techniques for preserving the privacy in order to conceal the identity of persons or groups are already employed in data science and often include anonymisation measures, access control and encryption of the data. Moreover, regulations on the limits to those practices and involving the user in the decision of what can be tracked and what data collected can be used and what not should be in place to protect them and simultaneously increase trust. The regulation in this area is currently a big issue already, growing at higher rates and hard to control, struggling to catch up with technical developments, and its enforcement probably presents one of the of the biggest practical difficulties.

Explainability and transparency would by itself not be enough but would presumably help increase the trust in algorithms. When a human can comprehend what exactly caused a

[5] "Ethics in artificial intelligence: introduction to the special issue ...." Accessed October 23, 2020. https://link.springer.com/article/10.1007/s10676-018-9450-z.

negative or unexpected outcome they are more likely to better cope with it than when simply confronted with an unexplainable negative response and no indication of a reason for it.
Nevertheless, possibly not all tasks require full transparency or full explainability, as there is a trade-off between optimal and transparent operation. The costs of implementing such additional functionality beside the main function, in case feasible, need to be evaluated and contrasted with the expected benefit of it. When the focus is solely on the full completion of the task itself, it might be irrelevant to understand every detail of its realization as opposed to those where to some extent should be explainable to its users and affected parties. Going back again to the example of a machine taking the decision of whether a person is eligible for a loan, as this person is directly affected by it, they should have the right to get an explanation.
Nevertheless, we need to keep in mind we are also not really able to fully explain how we make our own human decisions. In the best case, we come up with a plausible explanation why, mostly based on logic. This could indeed be the reason. But there's another possibility in which you believe what you say to be true, which means technically it is not a lie, but it is actually also wrong. This could for example occur when someone is asked why a certain person was or was not hired. This happens because we can never be aware of and control all variables involved in the decision making process. We are inevitably dependent on many factors, many of which we are not aware of. Therefore, we should not forget explainability and transparency are also unresolved issues about humans when asserting machines must definitely expose them.

One of the main aims of current social robotic research is to improve the robots' abilities to interact with humans. In order to achieve an interaction similar to that among humans, robots should be able to communicate in an intuitive and natural way and appropriately interpret human affects during social interactions. Similarly to how humans are able to recognize emotions in other humans, machines are capable of extracting information from the various ways humans convey emotions—including facial expression, speech, gesture or text—and using this information for improved human computer interaction.[6] Some tasks carried out by robots which imply direct interaction with humans could truly benefit from showing empathy towards them. To leverage these emotional capabilities by embedding them in humanoid robots is the foundation of the concept Affective Robots, which has the objective of making robots capable of sensing the user's current mood and personality traits that adapt their behavior in the most appropriate manner based on that. The ability of recognizing human traits to be able to accordingly adapt, as we humans try to do, is a definitely meaningful one, since it can pose considerable advantages and a much better execution of their job. Emotion recognition mechanisms and emotional intelligence is indispensable for them to understand us and work for/with us in areas such as assistance,

[6] "Affective Robots: Evaluation of Automatic Emotion ...." Accessed October 23, 2020. https://www.researchgate.net/publication/325688779_Affective_Robots_Evaluation_of_Automatic_Emotion_Recognition_Approaches_on_a_Humanoid_Robot_towards_Emotionally_Intelligent_Machines.

healthcare or companionship. Few people would argue it is not a very valuable feature for humans in the same position, which leads to the question of why it should be different if the same task were to be executed by a robot. Joseph Weizenbaum argued already in 1976 that we require authentic feelings of empathy from people in these positions.[7] If machines replaced them, we would find ourselves alienated, devalued and frustrated, for the artificially intelligent system would not be able to simulate empathy. His view at that time was that artificial intelligence, if used in this way, represented a threat to human dignity. On the other hand, having the robot "feel" (either truly, given feasible, or artificially coded) emotions in a similar way humans do, or having it express them might not be desirable. In any case, it should be agreed and fulfill a specific and beneficial purpose.

Exhibition of ethical behavior is the next major issue. The enterprise itself is genuinely complex, since there is no universal agreement on what "ethical" means to start with. There is agreement that ethical frameworks should be developed but it does not extend to how these should be implemented. What is clear is that this must be addressed properly in order to increase trust. This issue becomes very relevant and is highly discussed in some contexts where it plays a very serious role, as in those where human lives are at stake, e.g. in a controversial topic as self-driving cars. Machines making moral choices is key. Concretely, in the framework of self-driving cars, several ethical problems arise, namely who's responsible in case of a setback, safety, and the very well-known trolley problem applied to autonomous vehicles (AV), among others. As for liability, this remains a legal problem. As for safety for both the passengers and non-users, the trade-off between travel times, efficiency and safety, and ethical questions in case of an inevitable accident remain engineering problems. The trolley problem is a very well-known experiment in philosophy and psychology modelling an ethical dilemma, in which while driving a trolley car, the brakes fail, and on the track ahead are five people you'll run over. You can still steer onto another track, on which is one person who you will kill instead of the five: it poses the dilemma between unintentionally killing five people versus intentionally killing one. What should you do in such a situation? This issue is gaining back a lot of attention because of how it relates to AVs, and how there is the opinion an adaptation of this should be solved before their deployment in the roads. As a counter-argument, this should not apply since in the absence of own intentions from the cars the philosophical issue becomes irrelevant. Therefore, there would not be a strong direct connection in AVs. The situation is also different, as they do not drive on rails with predefined paths where you know for sure those people are going to get killed. On the roads there are many more variables, more uncertainty, and more possibilities to react. From an engineering perspective, such questions can be addressed more efficiently by focusing on preventing such occasions by design as much as possible. Focusing too much on theoretical situations where all variables and conditions are given and immutable draws attention away from something more important, which is building safer machines by design. Such a catastrophic brake failure might be easier to prevent, or detect with more time in advance in an autonomous vehicle

[7] Joseph Weizenbaum, quoted in McCorduck 2004, pp. 356, 374–376

fully equipped with sensors, so that the situation does not occur. Of course, this does by no means remove the urgent necessity of carefully evaluating what the minimisation of damages would be in the undesirable event of it being inevitable. Another point to consider is that based on its continuous learning, with successful and sufficient training AV will be continually improving their decision making process, at a much higher rate than any human driver. In the end, AVs efficiency and safety should not be compared to perfection, even if that would be alluring and a marvellous final goal, but to human drivers. They need to become on a first step at least as good as we currently are, not immediately perfect. The development and improvement of AVs from their 95% readiness to 100% readiness might take longer than the rest of the way, because that's what will make the difference in reliability and therefore also in trust. How likely is it that you're having an accident when you're on the driver's seat watching the car drive? Given it is a life-critical system, if the probability is not truly low, the decision will be they're not ready.

This is not the only situation where ethics comes into play. The problem of bias in AI has become very well-known in recent years. It has been shown how algorithms which showed apparent good behavior have actually been operating with background bias, such as discriminating against or not working for minorities e.g. in face recognition. These algorithms may be mirroring the bias in the data presented to them, of the organizational teams, the designers, the data scientists who implemented the models, the data engineers that gathered the data, etc, causing harmful results. This bias may be introduced via training data in a deep learning algorithm (perhaps because there is less training data for specific minorities), which can include previous biased human decisions as well as historic inequities and also by letting it freely interact with a biased world, for example examining and interacting with the internet. A solution to this issue is both indispensable and hard to achieve. Take into account we are requiring machines/robots/AI systems to get something right which we seem to not be capable of doing ourselves. Humans are so keen in having machines be ethical and so afraid of them not being it when the actual world is very far from generally behaving ethically. Everytime I hear scientists mention this topic I get a humble feeling, because, in my opinion, as soon as you reflect about this you become aware of our own imperfections and how these are undeniably shaping the ethics of the machines we're creating. We as a society are responsible for these biased decisions in algorithms that inadvertently discriminate by gender, race, or sexual orientation, among others, that end up hurting people. Of course we don't want them to copy us and repeat our mistakes. We should keep trying to find ways to prevent and correct bias in AI, in pursuit of fairness and models that have equal values across all groups. Diverse teams, including people of different characteristics, such as gender, race and background can surely help find it and mitigate it. Perhaps, also more than just technical approaches are going to be needed and instead a combination of multi-disciplinary experts. On top of that, regularly monitoring autonomous systems to detect issues quickly can also be of help. On the other hand, it might be worth it, in my humble opinion, to keep addressing the underlying issue in parallel. Moreover, we might even be able to take advantage of the several ways a bias-checking AI potentially could unveil those biases gone unnoticed to improve human

decision-making, since deep learning systems typically disregard variables that do not accurately predict outcomes. Thus, as a conclusion, humans and machines working together to mitigate bias seems to be the best option.

Once better social acceptance is given, some predict a progressive merge with machines in the following decades, with approaches going in the direction of enhancing human capabilities through transhumanism practices such as brain-computer interfaces.

Challenge 2: Automation and Employment

“Robots are stealing our jobs”, “What are you going to do when AI steals your job?”, “Will a robot take your job?”, “Wells Fargo Predicts That Robots Will Steal 200,000 Banking Jobs Within The Next 10 Years”, “Robots are stealing our jobs. Here's what happens when one steals mine” are real newspapers headlines from the last few years. There’s no shortage of media screaming about how robots and AI are stealing, destroying and taking jobs away from people. The subject surely makes for great headlines, and generates lively debate. Perhaps, a better mindset should however not include words like “steal” in it but rather “transformation”, or “change”. It is undeniable that the impact from automation, robots and AI in the labor market is a tremendous one. It’s already here, as we see those changes and we can expect many more to come. Even jobs that were thought as “irreplaceable” are looking different now just 5 or 10 years later, as under the current perspective this does not only affect low-skilled jobs but also analytical ones. If this happens, many jobs will very probably be lost to AI technology in the following years, and we might end up living in a world with mass unemployment of people who don’t have the set of skills required for the economy, so that humans will need to embrace the change, and find new occupations.
The implications are plenty and all of them profoundly complex. On the one hand, there are economical issues: How are displaced from the workplace people going to be able to financially survive? How are countries going to fund social benefits when the labor market is mainly composed of robots? On the other hand, we also face other kinds of social and emotional dilemmas: What are they going to employ their time on to fill that blank, and receive the social and mental benefits their job provided? How are they going to find meaning when they’re not needed by the society? Will this transformation increase wealth inequality among humans?
When everything is sustained by the idea that every human has one job, fulfills a need, gets paid for it, and pays taxes from which social benefits for all can be available, clearly everything falls into pieces when you eliminate the key element: humans have jobs.
The change is so profound that it absolutely can not happen overnight or otherwise we’re doomed. This should be a progressive change, not a sudden one. There’s a lot that needs to adapt, and it will take time, years or even decades. Therefore, in my opinion the true threat is the potential fast pace of change and not the change itself. States, corporations,

industries, and all social structures need to react quickly, and have adaptation plans ready, as the world is rapidly changing.
There's a limit for every social system to how many unemployed people they can have while keeping the system functioning. This limit should be kept in mind to set the boundaries and therefore regulate the amount and frequency of previously human jobs to get replaced by machines. That will limit how fast technology progresses, or rather gets deployed, and will help slow down the process to a pace the society is able to absorb it. As Yuval Noah Harri puts it in its book Homo Deus[8], dealing with this new social class economically, socially and politically will be a central challenge for humanity in the coming decades.
Besides that, still on the economical side, new ways need to be found or invented to fund the social benefits. Many advocate for taxes on robots, and the deployment of a Universal Basic Income (UBI). While this idea sounds promising, it will need to be carefully implemented and regulated, as my opinion is this won't be sufficient. The markets will change, the prices might possibly dramatically change and, although this might be on the right track, it won't be as simple as giving everyone some purchasing power, and more efforts will be needed.

With regards to the wellbeing and emotional side of the consequences, and what the people becoming unemployed would be doing, I believe an offer of life-long learning programs, as well as the necessary helps and tools for retraining and adaptation to new undertakings, many of them with high probability about tasks we can not even imagine today, will be necessary. So will also the flexibility on their side, and an open mind to what the future might bring in contrast to the expectation of a profession for life as it was decades ago.
A sort of continuously reinventing oneself to keep up with the times, and taking on new paths might be the best solution to keep being "in the system", keep having a meaningful occupation that satisfies their aspirations and helps achieving a feeling of accomplishment and belonging.
In such an environment of accelerating transformation, we will have to become polymaths to survive. Or, in the words of Charles Darwin: "It is not the strongest or the most intelligent who will survive but those who can best manage change."

Challenge 3: Regulation

There's agreement that regulation of AI practices are needed as the alternative responses such as doing nothing or banning development are both truly impractical. Academics debate the process of how governments could go about creating legislation for roboethics. A big part of the AI community, researches, scientists, as well as many other public figures argue that governments are not doing enough, and are really running behind. Regulators

[8] Harari, Yuval N. *Homo Deus: A Brief History of Tomorrow*. Signal, 2015.

are not anticipating but rather many years late, as they're still coping with understanding and regulating the internet, which has been here for some decades already. Actions should be taken fast instead of waiting until something happens to start thinking about it. AI is often seen as a science fiction topic, very far in the future, if anything, and not as urgent as other issues right here right now which need more attention. To me, it is essential to deal with it now and make the world realistically prepared for these developments. A balance needs to be found, for regulating too much would prevent development and the incredible gains, and regulating too little might have negative and perhaps even irreversible consequences.

What needs to be regulated? AI law and regulations can be divided into three main topics, all of them of high importance: governance of autonomous intelligence systems, responsibility and accountability for the systems, and privacy and safety issues. The most compelling aspects are presumably accountability and data privacy. Since these technologies rely on data, access to it must be allowed while at the same time being cautious and having strict privacy guidance. Privacy is a delicate issue which the public needs to understand. Every piece of data or combination of pieces that can be employed to create some sort of personally identifiable information about you and the way you live is personal data. A majority of consumers today are too complacent in this regard giving away precious information about their lives that might end up having negative consequences for them, if they land in the wrong hands. It is very tempting to abuse this information for people who share different values than we do. The same vulnerabilities and issues caused in the cybersecurity sphere, hacking and phishing we can expect to see in AI, as there is no reason to believe this will be treated differently. it is critical that there are safeguards and that companies are liable to protect the data adequately. Professional organisations should be held accountable for everything regarding the data they handle, which will be the motivation for them to show that they comply with the legislation and follow all necessary processes. A reason why this great endeavour is so complicated is the collapse of two forces: on the one side, the need of a fast time-to-market and on the other side a "clean" development. That's what makes introducing liability so crucial.

The EU announced in early 2020 their intention to create more access to data to be able to compete with the USA and China who currently present a more relaxed attitude towards the topic and, without compromising the rights of the citizens, they pursue the creation of a safe data sharing ecosystem. They want to incentivize companies collecting data to bring it together and share it for common good and propelling development made in Europe. The data should either not be not personal or be collected with consent and also data that does not compromise the security of the nations.
Furthermore, concerns about the potential abuse of facial recognition technology arise. China's mass surveillance of its citizens is a well-known example of it. The EU stepped away from a ban of the use of the technology in public areas and left it up to member states to decide.

Beyond that, also the regulators need to be regulated. Constraints on how governments use information should be in place by creating different legal standards so that they are allowed to use it only when necessary.

Ideally these regulations of AI that affects all of us would be taken globally but since this is very close to impracticable, a good start might be regulation at national level with dialog between nations regarding issues that connect them to look at them together, such as warfare. Perhaps in a few years the feasibility of more cooperation becomes true.

Although one may argue these regulations are taking too long to appear and they don't progress at a rapid enough pace in parallel to the development, it is true to say that some efforts in that direction have already started. To name an example, I will focus on the efforts of the European Union (EU), supported by the High-Level Expert Group on Artificial Intelligence[9], guided by a European Strategy on Artificial Intelligence, whose efforts started a couple of years ago. In 2019, the European Commission published its "Ethics Guidelines for Trustworthy Artificial Intelligence (AI)"[10] and its "Policy and investment recommendations for trustworthy Artificial Intelligence"[11]. The aim of the Guidelines is to promote *Trustworthy* AI. Trustworthy AI, as they describe it, has three components, which should be met throughout the system's entire life cycle:

1. It should be lawful, complying with all applicable laws and regulations
2. It should be ethical, ensuring adherence to ethical principles and values and
3. It should be robust, ensuring AI systems will not cause any unintentional harm.

It is divided in three chapters. The first one sets the foundation of Trustworthy AI by laying out its fundamental-rights based approach. It describes the following ethical principles: respect for human autonomy, prevention of harm, fairness and explicability that must be adhered. The second chapter focuses on realising Trustworhty AI, which translates the principles into key requirements that should be implemented, namely human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being and accountability. Finally chapter sets out a (non-exhaustive) assessment list to operationalise the requirements addressed to developers and deployers of AI systems.
Their Policy and investment recommendations for trustworthy Artificial Intelligence is the next step and presents a set of policy and investment recommendations on how Trustworthy AI can actually be developed, deployed, fostered and scaled in Europe, all the while maximising its benefits whilst minimising and preventing its risks. For the purpose of

---

[9] "High-Level Expert Group on Artificial Intelligence | Shaping ...." Accessed October 23, 2020. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.
[10] "Ethics guidelines for trustworthy AI | Shaping Europe's digital ...." Accessed October 23, 2020. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[11] "Policy and investment recommendations for trustworthy ...." Accessed October 23, 2020. https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

contributing to individual and societal well-being, they formulated 33 concrete recommendations addressed to the European Institutions and Member States.
In 2020, they published a White Paper: “A European approach to excellence and trust”[12], consisting of two main parts, an ‘ecosystem of excellence’ and an ‘ecosystem of trust’. The latter outlines the EU's approach for a regulatory framework for AI. In its proposed approach, the Commission differentiates between 'high-risk' and 'non-high-risk' AI applications. Only the former should be in the scope of a future EU regulatory framework.

Challenge 4: Common sense

Today’s AI systems are getting impressingly good at detecting pedestrians but they actually do not know what a pedestrian is. In image recognition tasks, they accomplish their goals with remarkable accuracy in many cases, while some other times making very weird mistakes, utterly different from the mistakes humans do. For example, there are algorithms to detect animals in images by recognizing patterns. After being trained with thousands or millions of images they are able to tell them apart in almost every case but, the thing is, sometimes for the “wrong” reasons. Some well-known examples of these weird mistakes prove that putting the same animals with different backgrounds or in settings where they are rarely found on completely confuses the system causing a mistaken outcome. Sure, humans might find a picture of cows at a beach somewhat random but no one would have a problem in recognizing they are, indeed, cows! The same way, animals that usually live in a snowy environment have been able to trick the systems when depicted out of it. For deep learning algorithms that recognize patterns, this is a challenge. The fact that the algorithms themselves decide what’s relevant in order to recognize a specific thing makes it really hard. What’s missing is something much more difficult to teach or train. Something we don’t need to do with humans since we develop it inadvertently early enough in our first years by observing the world: common sense. We build models of the reality we live in.
Can we call those systems intelligent when they don’t know what a pedestrian is? Or a cow?
Our common sense helps us, as Yann LeCun puts it, to fill in the blanks. We are able to infer the state of the world from partial information and uncertainty. We are able to infer the future and past events, and fill the missing segments in images and speech. This ability to infer and predict is the essence of intelligence because it allows us to reason, understand and answer complex questions. We form models of the world since we’re babies. That’s how we learn most of what we learn through life and this is what machines are missing. In terms of machine learning, it would be resembled by “unsupervised learning”, in its essence different from supervised and reinforcement learning, more common practices. It makes

[12] "White Paper on Artificial Intelligence: a European approach to ...." Accessed October 23, 2020. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

no use of any labeled data or any reward function. Instead, the algorithm is let to explore on its own to discover unknown patterns.

A total integration of robots in the society will not take place, in my opinion, until they acquire a common sense, which allows them to live among humans and interact with them at every level. Now, finding out how to implement that, still remains in our to-do list.

Challenge 5: Ethics and the existential risk

Ethics in AI and robotics raises plenty of concerns of various sorts. In fact, as seen over and over again throughout the centuries, this is a typical response to all new technologies during their appearance.
With time, some of those concerns and predicitions turn to be completely irrelevant and forgotten, some are proven to having been fundamentally wrong, some are generally correct but only moderately relevant and others are broadly correct and fully relevant. What will be the case for the current concerns regarding AI and robots?

A positive outcome for superintelligence could preserve Earth-originating intelligent life, contribute to areas such as mitigating disease, poverty and environmental destruction, and could help us “enhance” ourselves and fulfill our potential. Many state that building AGI would be the biggest accomplishment of humanity, not comparable to any other achievement so far. But is it possible to develop a superintelligence? We humans are presumably not the peak of possible intelligence, and thus rather much higher levels are theoretically possible. Intelligence is a matter of perceiving information, processing and applying it. Therefore, as long as we continue experimenting and making progress exploring new ways of information processing, we (humans) might eventually build AGI at some point. A lot of researchers in leading AI labs, such as DeepMind, are already focused on doing precisely that. The problem here is, if we build machines that exceed us in intelligence or even are just as smart as us but with a million times faster “thinking” (computation), how could we constrain them?

In “Artificial Intelligence: A Modern Approach”, the authors assess that superintelligence "might mean the end of the human race". It states: "Almost any technology has the potential to cause harm in the wrong hands, but with superintelligence, we have the new problem that the wrong hands might belong to the technology itself." Researchers normally refer to it as “singularity” or “intelligence explosion”. The thesis that AI poses an existential risk (that could result in human extinction), and that this risk needs much more attention than it currently gets, has been endorsed by many public figures. Some of the most famous are Elon Musk, Bill Gates, and Stephen Hawking. The most notable AI researchers to endorse the thesis are Stuart J. Russell and I.J. Good.

As mentioned earlier, this is something it is being worked on right now by many groups in parallel and its progress is surely being observed very closely by many. Given the stakes, almost that to win this race is to win the world (provided you don't unintentionally destroy it first), this generates an ecosystem somewhat resembling a race against all others to get there first, and that's why we can assume we will continue to try to improve our intelligent machines. Elon Musk mentioned in an interview that the least scary future he can think of is one where we've democratised AI because, if one company owns it, it would be an immortal dictator from whom we can never escape.
To create the conditions to develop AGI in a safe way before all else would be ideal but given the aforementioned situation, it is likely that whatever way is easier to go will be the one chosen. This poses us in utterly alarming circumstances, as even mere rumors of someone getting close to develop it could cause disastrous consequences. Given that, out of responsibility, safety measures concerning the behavior of humans as they design, construct, use and treat artificially intelligent agents need to be urgently taken when doing research in AI. I would divide the necessary measures in the following three categories:

- Increase awareness in the pursuit of global cooperation
- Encouraging more work on risk-decreasing development
- Superintelligence control measures development

What exactly should we ultimately fear? Science fiction repeatedly depicts an AI turning evil and developing malicious objectives but, as it turns out, an AI suddenly turning malevolent is not the biggest worry or statistically probable outcome but rather superintelligent out of control agents that are too good in pursuing poorly specified goals. As Steve Omohundro, Nick Bostrom, and others have pointed out, the combination of a value misalignment with increasingly competent decision-making systems can lead to immense problems, perhaps even capable of inducing the end of our species. Many experts argue machines not sharing our goals is the most likely scenario, but how to ensure they share our goals for the future we want to create? And how can humans stay in control? The AI control problem is well-known both in AI and in philosophy and deals with the issue of how to build a superintelligent agent that will aid its creators while at the same time avoid inadvertently harm them. Its study is nowadays crucial, as it is motivated by the perception that humans need to solve the AI control problem *before* any superintelligence is created. In contrast to the existing (weak) AI systems, which can be permanently monitored and easily be shut down or modified when they misbehave, a true superintelligence, which is by definition smarter in solving all sorts of problems in the course of pursuing its goals, would easily become aware of the dangers of being shut down to accomplish its goals and thus interfere and do everything that needs to be done to prevent modification and shutdown.

A number of organizations are working in a technical theory of AI goal-system alignment with human values, share big part of their values and goals and are making remarkable contributions. Among these are the Machine Intelligence Research Institute, the Future of

Humanity Institute, the Center for Human-Compatible Artificial Intelligence[13], and the Future of Life Institute.
The Center for Human-Compatible Artificial Intelligence goal, for example, is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research towards provably beneficial systems. [14]
The Future of Life Instititute is a non-profit research institute and outreach organization that has as mission "to catalyze and support research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course considering new technologies and challenges."[15]
They propose to follow a total of 23 principles, known as the Asilomar AI principles, which under my perspective englobe all mentioned concerns in the previous sections with regards to safety, cooperation, responsibility, privacy, value alignment. The following are, under my subjective perspective, the most outstanding ones:

1) Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

11) Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

20) Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.

21) Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

23) Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.

Moreover, the Partnership on AI (PAI)[16] is a multistakeholder organization that brings together academics, researchers, civil society organizations, companies building and utilizing AI technology, and other groups working to better understand AI's impacts. The Partnership was established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society. The Partnership was formally established in late 2016, led by a group of AI researchers representing six of the world's largest technology companies: Apple, Amazon, DeepMind and Google, Facebook, IBM, and Microsoft. In 2020 it represents a community of 100+ partner organizations. According to their website, their thematic pillars are:

- Safety-critical AI

[13] "Center for Human-Compatible AI." Accessed October 23, 2020. https://humancompatible.ai/.
[14] "Center for Human-Compatible AI." Accessed October 23, 2020. https://humancompatible.ai/about#mission.
[15] "Future of Life Institute." Accessed October 23, 2020. https://futureoflife.org/.
[16] "Partnership on AI." Accessed October 23, 2020. https://www.partnershiponai.org/.

- Fair, transparent and accountable AI
- AI, labor and the economy
- Collaborations between people and AI systems
- Social and societal influences of AI
- AI and social good

One of the most common reasons we are told to not worry about the rise of AI is the time horizon, i.e. it is far away in the future and therefore not necessary to think about it now what, to me, sounds like a very weak reason. Should it happen or not, and should it happen sooner or later, my opinion is anticipation is key. However, a big part of the society, both as individuals and as nations seem unable to produce a clear reaction to the possibility of it happening.
Given the stakes, namely no less than the continuity of our species, I deem crucial to prioritize risk-reducing strategies such as progress in the AI control problem over risk-taking strategies in AI development. These risk-reducing strategies should attempt to answer the following question: what types of safeguards, algorithms, or architectures can programmers implement to maximize the probability that their recursively-improving AI would behave in a human-friendly, rather than destructive, manner after it reaches superintelligence? Nick Bostrom recommended the altruistic global adoption of a common good principle: "Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals". Others phrase it with different words but towards more or less the same target, such as “maximize freedom of action of humanity”, or “help humans flourish”.[17]

Isaac Asimov’s three laws of robotics are a widely known set of rules introduced in 1942 and although they appeared in fiction they have had considerable impact on the topic of ethics in AI:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.[18]

These were intended as safety measures that can not be bypassed and must be incorporated in all fiction robots appearing in Asimov’s series. Much of his work was then spent testing the boundaries of his three laws to see where they would break down, or where they would create paradoxical or unanticipated behavior. His work suggests that no

---

[17]Luke Muehlhauser and Nick Bostrom (2014). Why we need friendly AI . Think, 13, pp 41-47 doi:10.1017/S1477175613000316

[18]Asimov, Isaac (1950). "Runaround". I, Robot (The Isaac Asimov Collection ed.). New York City: Doubleday. p. 40. ISBN 978-0-385-42304-5.

set of fixed laws can sufficiently anticipate all possible circumstances. The original laws have been later on altered and elaborated on by Asimov and other authors. Asimov himself made slight modifications to further develop how robots would interact with humans and each other. Asimov also added a fourth, or zeroth law, to precede the others:

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Asimov himself said in an article in 1981[19] the laws were obvious and he just managed to be the first to put them together, as well as that they can be applied to anything else. In the same article he mentioned he had an answer ready whenever someone asked if he thought that the Three Laws of Robotics will actually be used to govern the behavior of robots, once they become versatile and flexible enough to be able to choose among different courses of behavior and that his answer was, "Yes, the Three Laws are the only way in which rational human beings can deal with robots—or with anything else — but when I say that, I always remember (sadly) that human beings are not always rational."

An adaptation of those rules, derived substitutes of them to something we can work with, need to be introduced to be able to specify good behavior in such terms as "do X in such a way that no harmful consequences happen to humans".

In his book "Human Compatible: Artificial Intelligence and the Problem of Control" (2019), Stuart J. Russell asserts that the risk to humanity from advanced artificial intelligence (AI) is a serious concern despite the uncertainty surrounding future progress in AI. It also proposes an approach to the AI control problem. The main points of his approach are the following:

- Altruism. The machine should maximise objectives of the humans and not their own
- Initially they must be unaware of what those objectives are
- Through observation of human choices, they should learn about how we prefer our life to be, inferring our values from using a form of Inverse Reinforcement Learning

During a speech, he made a very clear example with the sentence "you can't fetch the coffee if you're dead". To understand that, the context needs to be explained, namely a robot whose goal is to fetch coffee might come to the idea of eliminating humans interfering with its task of fetching coffee, since they might try to switch it off before accomplishing the task. This short but powerful sentence illustrates perfectly the potentially fatal underlying issue.

My impression is, we grasp the potential of AI, at the same time we also fear it, and we don't know exactly what we want it for. Therefore, I think the human-compatible AI approach from Russell fits perfectly and should be used as a guidance.

Furthermore, as many academics point out, I believe that attempts to teach robots how to behave "in a moral way" will likely help make progress in understanding and improving human ethics as, for instance, suggested in "Moral Machines: Teaching Robots Right from

[19]Asimov, Isaac (1981). "The Three Laws". *Compute!*. p. 18.

Wrong”[20] from Wendell Wallach and Colin Allen, which would indeed be a very positive side effect.

Whatever may happen in the future, looking apart and doing nothing about it pretending AI and robots are not here is probably the worst we can do, both as individuals and as a society. They have huge potential to enhance our society in every single way. Let’s be smart ourselves and find out how!

[20] "Wendell Wallach and Colin Allen: moral machines: teaching ...." Accessed October 23, 2020. https://link.springer.com/article/10.1007/s10676-010-9239-1.