Ethics, Law and AI

Federico L.G. Faroldi

v. 1 – December 2022

Please check the new version of this document again at the end of the course in mid January

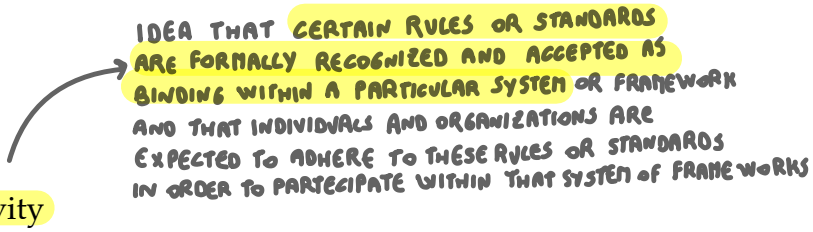
~~Crossed out~~ parts are not required. You can still read them of course if you are interested.

Table of Contents

Note to the reader

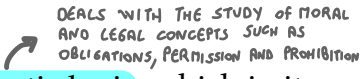
This draft is based on lectures delivered in the course "Ethics, Law and AI" at the University of Pavia in 2022-23.
Some chapters are transcriptions: as they follow natural speech patterns, there are sentences with some grammatical mistakes. In the interest of timeliness, these lecture notes are made available as is. Apologies in advance for typos and mistakes.
You can read ~~crossed-out sections~~ for context and for your personal interest, but they are **not required** reading. Make sure to check again these lecture notes by the end of the course.
Many thanks to Dr. Nathan Wood for help with some of the transcriptions, and to some students for making some of their notes available.
You are welcome to point out (serious) mistakes to federico.faroldi@unipv.it

Introduction



Formal Normativity

That there are genuine questions emerging from the use of formal tools in the investigation of normative phenomena is what the relatively new discipline of formal ethics, which has an unfortunate name and should instead be called formal normativity, argues.



This discipline has its roots in deontic logic, which in its modern guise came into being in the 1950s, and was inspired by formal epistemology, that is, a discipline that uses formal methods, such as probability theory, logic, and decision theory, to study traditional epistemological topics, such as knowledge, belief, etc.

What is the method of formal ethics? It is not easy to say. In fact, I know of no explicit reflections on the subject.
First of all, one must ask whether the discipline is descriptive or prescriptive.
In the former case, and simplifying by assuming that it adopts a scientific/abductive method, one must ask what are the data on which it is based. And this certainly depends on the metanormative position of reference.

For example, in law one might refer to laws placed, or to court pronouncements, in morality to intuitions, and so on.
In the second case, that is, when it is at least partly a prescriptive discipline, one has to start almost from scratch methodologically.

What are the advantages of formalization?

Increased precision by decreasing the vagueness of natural language. It becomes easier to make objections and see the consequences of certain premises.
Computability, at least in some cases, which is a basis for automation.
Automation has advantages and disadvantages. The advantages are speed, accuracy, and avoiding human input. This in turn has advantages, in terms of saving energy, decreasing errors, and discrimination, but also disadvantages, such as the impact on the workforce, the structure of society, and the introduction of other types of bias, such as algorithmic bias.

What are the disadvantages or costs of formalization?

Like any kind of modeling, formalization forces a focus on some aspects to the detriment of others, and thus can introduce biases and oversimplifications.
Moreover, the limit of what can be captured is restricted to the expressive power of the tools one uses.

Formalization, automation, and the new frontiers of artificial intelligence between ethics and law

As can be seen, the development of artificial intelligence is now unstoppable, or at least unstoppable. It is therefore normal to also see applications of artificial intelligence to normative issues, both ethical and legal. The chronicle is full of sometimes problematic situations caused by the use of artificial intelligence techniques in normatively relevant matters in ways that are often flawed and sometimes extremely harmful.

I would just like to mention the famous COMPAS system that has been used by some courts in the United States to calculate the risk of recidivism and thus determine the length of sentence for offenders. The problem is that after a few years of use it was discovered that this system suffered from a racial bias: specifically it predicted and then was shown falsely that poor and black people were more likely to commit crimes again and thus were sentenced to longer sentences.
Two problems: first this was shown not to be true, second the algorithm for determining the likelihood of recidivism was proprietary and therefore not accessible to either the parties or the judge. In this case we are still talking about a system to help a human person but think that many decisions relevant to our daily lives are or will be made by algorithms: opening a mortgage, being accepted university, being selected at a job interview and so on.

Recently an artificial intelligence system called Delphi has also been created that interacts with the user by providing ethical judgments.
For example you can ask this system what you should do in a particular situation involving people goods etc. from a moral point of view is it right or wrong to give twins the same games? This system is based on a data base of situations evaluated by human beings particularly in the United States and in the 21st century. As you understand this is a serious problem. The productions made by this system in fact reflect not morality or ethics which are prescriptive disciplines: they tell us what is right to do, but they reflect descriptive predictions tell how the average 21st century American interprets situations. So potentially assume the same system but ask is based on the views of the average mid-19th century American: in this case the system would respond that slavery and trivially right reflects majority opinion. But this is not ethics is not right wrong, prescription, but description of what the majority thinks.

In any case, this process has begun, and so it needs all our attention, in an interdisciplinary way, to be best managed and directed. In particular, technology should not be chased, putting a patch on problems that have already been created, but there is a need for the legal system and society to direct and guide

technological development toward acceptable outcomes and directions from a prevention perspective.

Moreover, since a computer approach to normative issues is inevitable, the study of logic and legal informatics becomes of primary importance: firstly in the study of the formalization of relevant notions such as that of obligation, permission, validity, value; secondly in the study of the formalization of normative reasoning.

→ WHERE DO WE WANT TO GO?

Quo volumus vadere?

Once we have established that these new tools are there, that they pose ethical and legal questions, and we have identified some tools that facilitate us in analyzing the problems we face, we have not yet answered the question: where do we want to go?

Specifically, we need to think, as a community, hopefully not local but global, far in advance about what kind of society we want to have and leave to our great-grandchildren.

Indeed, the issue of future generations and their rights is pressing but much underestimated.

It is we, being at a turning point in history, who must set a direction for technological development that is compatible with the needs and values of humanity and, hopefully, the rest of the moral patients.

It is at this point that we return to a fundamental problem in human reflection, simply formulated in a contemporary context: what is right to do, as an individual and as a society?

But you will not find an answer to that question in this course. Each of us must answer it to ourselves and to others.

The Debate on AI and Intelligence

The Turing Test

Turing, in a seminal paper published in *Mind* in 1950, titled "Computing Machinery and Intelligence", proposed to replace the question 'Can machines think?' with the Imitation Game, where there are three players: two humans and machine. One human has to guess, by posing questions in such a way as not to see or hear the other players, who is the other human, who is set out to help them, and who is the machine, which is set out to fool them to believe that the machine itself is the human.

This test replaces a psychological or cognitive understanding of thought, and more broadly intelligence, with an empirical or behavioristic understanding. It would be sufficient that an average interrogator were fooled for so many time (and empirical question) for something that behaves like a human than thinks like a human.

This behaviorist sensitivity is of course partly a by-product of the spirit of the time, coupled, however, with what seems a slight rebellion to the Oxonian ordinary language philosophy.

Some problems with the Turing's test are that the test might be chauvinistic: it only recognizes intelligence in things that are able to sustain a conversation with us humans.
Moreover, some have thought that the Turing Test is not sufficiently demanding: success in Turing's Imitation Game might come for reasons other than the possession of intelligence.

Or too demanding: there may well be features of human cognition that are particularly hard to simulate, but that are not in any sense essential for intelligence

It is also unclear the reason of the restriction of the discussion to the case of "digital computers": if the test that Turing proposes is a good one, then it will be a

good test for any kinds of entities, including, for example, animals, aliens, and analog computers.

Searle's Chinese Room Argument

It's historically the basis for the distinction between Strong AI and Weak AI: Weak AI might be able to imitate humans and pass the Turing test, whereas Strong AI has something Searle understands as consciousness, intentionality, understanding. Searle seems to be quite chevalier about which one he means.
In order to argue that one can imitate human behavior very well, perhaps in one of its highest peak, that of understanding stories and be able to answer questions, without having understanding, Searle constructs an argument based on a thought experiment, famously called the Chinese Room argument.
Searle imagines himself alone in a room. He has a set of instructions (a programme), in English, which he understands, on how to manipulate and respond to sets of Chinese characters, which he does not understand in any way. Having understood the instructions, he is passed phrases of Chinese characters through the slit beneath the door. Using the instructions, he responds correctly. Those outside the room, passing him the phrases, are wrongly lead to believe that there is a Chinese speaker in the room.

Searle's Chinese Room Argument aims to show that the Turing Test is inadequate because imitation is not sufficient for understanding, and understanding is not present in weak AI.

A contemporary definition of Artificial Intelligence

1) 'The study and design of intelligent agents, where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success (Russell-Norvig)

As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect.

Weak AI: programs that do not experience consciousness or do not have a mind in the same sense people do, but can (only) act like it thinks and has a mind and consciousness

Strong AI: ability of an intelligent agent to understand, feel or think like a human

Narrow AI: ability of an intelligent agent to learn and perform a specific task, often with at least human proficiency

General AI: ability of an intelligent agent to learn and perform any intellectual task that a human being can

Superintelligence: a hypothetical agent that would possess intelligence far surpassing that of the brightest and most gifted human mind.

singularity: a hypothetical point in time at which technological growth becomes uncontrollable and irreversible, resulting in unforeseeable changes to human civilization

Obviously not equivalent.

Introduction to Ethics

This section is about ethics. Since this is not a book about ethics for philosophers, it's going to be really a bit of a naive introduction. But I hope that the main divisions, sub disciplines and approaches and topics will be clear.
On the other hand, when it comes to ethics, we start from a particularly good point, because contrary to other abstruse philosophical disciplines, perhaps metaphysics, perhaps philosophy of logic, and stuff like that, we do think and deal with ethical quandaries, if not full blown moral dilemmas, since we are little kids. And why is that? Well, because it is almost natural, I mean, extremely, extremely common for human beings to ask themselves: what should I do? and to tell other people in their community: you should do this or that.
The 'should' in 'What you should do' is not really an instrumental should, like: "if you're hot, you should turn on the airco. If you're cold, you should wear a coat". The latter are *instrumental obligations*, and in general they are not really part of ethics. So a terminological distinction to start with: for the purposes of this course, we do *not make a distinction between ethics and morality*. In theory, there's an etymological distinction, as you might be aware of: ethics comes originally from ancient Greek *ethos*. And morality, comes originally from Latin *mos, moris*, but these two terms do have vague meaning in common because they both mean habit, or custom. They have a sort of normative connotation. There have been theories in the history of philosophical thought, that have made the great deal about a distinction between ethics, which would be like superior because it's about the collectivity, about society, and morality, which is still important but it is about the individual. So we won't make this distinction.
And so if one wanted really to have a *very loose definition of ethics*, one could say that it's about *what one should do*, morally speaking, or *all things considered*.
So you might have competing obligations, competing interests, competing goals. But ethics also claims a preeminent role. And this is partly recognized also, for instance, by parts of the law. Sometimes this goes under the label of *conscientious objection* in countries whose legislation admits of such a thing. So you might object to military service based on your morals or if you're a doctor, you can object to perform abortions based on your moral convictions.

Okay, that's it for the for this terminological parenthesis, although there will be more as we explore more issues.

Ethics is concerned with what people ought to do or shouldn't do, or perhaps must do. Or, in fact, what is permissible, morally permissible for them to do. So it's about your duties, but also your *permissions*. So one might say that abortion is permitted. Ethically speaking, that doesn't mean that it's obligatory so that everyone has to perform abortions, right? Or everyone who's pregnant has to abort.

But there's also and we touched a little bit upon it already. It's also customary to proceed in a dichotomic fashion. So first, we should distinguish between what it's considered, descriptive or "descriptive ethics". And what it's considered a normative endeavor, or "normative ethics". So what's the descriptive endeavor? The descriptive endeavor is perhaps "proper science" in the sense of an empirical science, which might take different forms, of course, the idea being that you just investigate: What was the ethics? Or is the ethics of a particular population? In a certain timeframe? So what are the ethical beliefs or behaviors? Or in general, what's the ethics of the Roman upper-classes in the first century after Christ? Or of the East Coasters in the United States in the first quarter of the 20th century? Or perhaps, what are the ethical beliefs of contemporary Americans?

This is not a normative endeavor. It might be normatively connotated. But only to the extent that such normative connotations also pervades other sorts of descriptive research. So this could be a *historical* endeavor, or it can be a *sociological* endeavor. But per se, it's not an ethical and legal. Another version of descriptive ethics, or at least that's how I like to think about it is those who investigate *how morality emerged*. So are humans the only ones to have what we call morality? Or what's the status of animal "moralities"? In other primates, for instance, what are the biological mechanisms that favor the emergence of moral thought? And what are instead the social requirements or events that led to the emergence of morality? There are fascinating competing hypotheses about this. But also this is not a normative endeavor because it just describes what certain people at certain times believed or not have acted in accordance to. So when it comes to normative ethics, then it's really about the 'must' or the 'should'. That's what makes a difference. Of course, on the other hand, you could, you could also use descriptive

statements to describe what their obligations are without really endorsing them. So there's a whole field of research about *descriptive versus normative statements*, and what's their semantics, what's their logic, which, obviously, we cannot really get into them. In this short amount of space we have, we can say that normative ethics is a normative enterprise as it also needs to be separated from at least two other disciplines that are sub disciplines in this case are at least pursued still within philosophy departments? If that's any good indication of the fact that they're sort of within ethics rather than sociology or history or stuff like that. One is called *meta ethics*. And one is called *applied ethics*.

So really, in this sense, when, especially when it's contrasted or compared to meta ethics, normative ethics is also called *first-order moral theory*. Why? Well, because the idea is that it will give you a concrete recipe on what to do on concrete issues, not muse about ethical issues themselves. So it really tells you okay, here's the situation. *That's how you should act*. Of course, there's also theorizing about this. And in fact, we can distinguish a few major first order moral theories that range from the *deontology, to consequentialism, to virtue ethics, to particularism*. And we'll see each of these theories in turn in a little while.

Meta ethics, instead, as the name itself suggests, is *the study of questions about ethics*. And these themselves can be grouped, or at least, we can distinguish different levels or different plans of study. So there's an *epistemological level*: how do we get to attain knowledge in the ethical domain? How do we know what it's obligatory for us to do? What is our duty? Or how do we know what's the correct ethical theory? Is this knowledge innate? Do we construct through a collective process? Is it based on desires? So it's pretty subjective and so on, so first epistemological, second, *the study of the language of ethics*. So what does 'should' mean? Or what does 'may' mean? *What's the meaning of 'good'?* Right we should pursue the goods or lying is bad, what's the meaning of bad here? How do they this notions interact? From our linguistic point of view, how do you distinguish the evaluative uses? May and good this kind of words from their descriptive uses. What's their *semantics*? What's their *logic*, how do you get to conclude certain things in ethical reasoning, if you don't know whether normative language and that follows the same logic, that it's followed when it comes to prescriptive discourse. So, language

and logic level and last but not last, *the ontological or metaphysical level*. So, do these two ethical concepts or notions exist? Do
moral facts exist?
Is ethic, ethics subjective, just in our minds or objective out there in the world, there's a social construction. So, *meta ethics is the study of ethical questions or second order moral theory*. And we can sort of divide it in three: the epistemological domain, the linguistic domain and the ontological domain.

And then there's Applied Ethics. *Applied ethics is just the application of certain first order moral theories to specific concrete issues.* So, ethical issues in business, for instance, or issues in healthcare or issues in profession, like if you're a lawyer, or if you're a politician, you have like a *professional code of conduct* that also takes into account obligation towards your clients, obligation towards your colleagues or the electorate and so on. So, we introduced ethics. And we distinguished a descriptive approach to ethics, a normative approach to ethics and within a normative approach to ethics, we distinguished between meta ethics, normative ethics, in a strict sense, also for like, first order more theory, and applied ethics. And we saw that mathematics can be split into an analysis, epistemological, domain, epistemic linguistic domain and the ontological model.

Normative Ethics

We are now going to see a few *theories in normative ethics*, in the strict sense. When it comes to normative ethics, there are several theories as you might expect. As humans, we basically almost never agree on anything. And this is also an issue in ethics called *moral disagreement*, which might we might see later.
There are different ethical theories, and they all trying to answer in somewhat general terms (otherwise they wouldn't really be theories) to the question: *What should we do? What is the right thing to do? What is just what is good?*
This is the general category of questions that theory is trying to answer. But of course, since they are different theories, they also give different answers. And sometimes ethical theories are kind of tied up to certain ontological or epistemological theories. And so they also have their own views on what the right notion when it comes to what ethics is. So someone might say: What should we

do? Or what's the right thing to do? Is not the right question to answer. A direct question to answer is, what is good? And how can we realize what is good? But anyway, that's a different level of detail that we're not interested in in this context. So just to have a complete picture, we're going to briefly introduce and talk about four main families of first order moral theories deontology, consequentialism, virtue ethics, and particularism.

1) So let's start with deontology, which to the question 'What should we do?' replies with 'our duty'. That's the root of the word itself from *deon* in Greek, it's related to what should be done our duty. From which also other words derive, such as *deontology* or *deontic* logic, and so on. And then, so the idea is that, *there is a principle that you should follow*. And that's what encodes our duty. Of course, this principle or principles, it's just, this is a bit like just the general format of the theory, because when I tell you, Well, you should follow this principle, do your duty, well, now you don't really know what to do either, because what's your duty? Right? So then, of course, each deontological theory (because as you can imagine, there are very many deontologist theories) will tell you what the duty in question is, what the principles that you have to follow are. So in this sense, these theories that we're talking about now are first order moral theories, because they're really telling you what your duty is, right? Or how you should in fact, act in concrete situation, or what should I do in this particular situation? Well, you know, I should follow these principles. And so, then I should also follow these principles in this particular situation. So to give you like a crude, very, very crude example of a theory that, of course, underneath them, we could call deontologist theory is the kind of morality that emerges from the Old Testament. So if you think about the *10 commandments*, which have been sort of, you know, considered absolute moral
principles. They really give you like, a conduct of action, what do you do? Well, you should
respect your parents, you should not steal, you should not commit murder, you should not, you know, and so on and so forth. So, if someone comes in your door, and says, Oh, can you please hide me because there's a murderer who was following me, and trying to murder me. And then you had this person, and then another person rings your doorbell? And asks, whether you've seen such and such

person, in fact, hiding? And the question is, okay, what should I do? Should I tell him? Can I tell this person was probably a murderer, was looking for this other person that I'm hiding it and where it is? Or should I lie? Well, so you should check against this list of principles. And is there a principle that prohibits you to lie? Well, if there is, then you shouldn't lie. To do the right thing. If there isn't, well, you can you can basically do whatever you want. If there are other principles, or other principles are not interpreted, interpreted in such a way as to suggest that if you were not to lie, this would cause a murder and also avoid you know, murder. It's not just not murder, murder yourself. As you can already see, one perhaps could say essential trait of moral theories is that they have to deal with what are called *moral dilemmas*. Suppose that there's a principle that says you should preserve human life and avoid people being murdered, let's say, and the principle that tells you shouldn't lie, then in this particular situation, it's really unclear what you should be doing, because if you don't lie, the person probably gets murdered. And so, you have respected the second principle, and you've contravened the first. And if you lie, you kind of preserve or obey the first principle, but you contravene the second. So, there are different ways this dilemma could be solved for instance, there could be *a hierarchy of principles*. So, one could say that the principle 'Do not murder' or, you know, 'preserve human life' has *greater importance* than the principle do not lie. So, contravening the principle, do not lie is justified if it's guaranteed to preserve the human life. About one could say that, as long as there are principles, there's not one principle which is more important than others and so none of them should be contravened. Okay, this was an example of a deontological theory that has more than one principle or more than one duty. That has to be followed, if you want to do the right thing. But they are the logical theories whose *theory of value*, and it's called technically, *monist, so where the principle is only one*. And *Immanuel Kant*, the German, perhaps now it would be Russian. philosopher, is perhaps the best known theorists. And now we can't really get into the details of contents theory. But there's a few different formulations of this *categorical imperative*, such as "always act in such a way that the maximum of your action can be a principle of universal legislation". I mean, they're all slight variations. And I hope that no ethicists is reading these lines. But they're simplified to facilitate understanding, all variations of the so called *golden rule* that you find in so many, so many cultures

around the world, not just in Western cultures: "Do unto others what you want others to do to you", or words to that effect. I mean, there are tons of reasons why, in fact, Kant's category categorical imperative, and the golden rule are totally different. But I just wanted to convey you the idea that *there's one principle that tells you what your duty is to follow this principle*. And that's what the principle says, of course, in a sense. This principle *does not contain concrete duties*. So it doesn't really tell you, you should not lie or where you should not kill, and so on and so forth. So it's for you by using your rational powers in the case of count that in that particular situation, should figure out how to apply this general principle to the action state.

Another deontological theory, *pluralist* and more contemporary, is the one theorized by William David Ross, an English philosopher who lived and worked in a 20th century. There are multiple duties: the duty of beneficence, the duty of non maleficence, and so on. These duties are however, as we say, prima facie. So it's true that there is a bit of a hierarchy between those principles. But in the concrete, in a concrete particular situation, you should figure out what your actual duty is. Which sometimes it's not the same as the most important of these four principles in theory. What you have to do, all things considered, is your actual, or all-things-considered, duty. Okay, so the deontological theories tell you that you should do your duty. That's how you should be acting and your duty consists in following one or more principles, perhaps complicated by further operations to decide what's your actual duty in a particular situation, especially when it comes to solving dilemmas. So that's one very big family of first order moral theories.

Consequentialism

But there's another very big family of first order moral theories that it's broadly labeled consequentialism. Consequentialists tell you that what you should do in a particular situation does not depend on your action or actions conforming to a particular principle, but it depends on the *consequences of your action* or actions. So that's what your duty is. But then *how do you evaluate the consequences of your action?* What should the consequences of your actions tend towards. And here, there are, of course, different theories. Different answers that are different are

given by different consequentialist moral theories the haps, the original historically important version of consequentialism was *hedonism*. So, actions are evaluated by the amount of pleasure, they cause, or the amount of pain they avoid. This was later corrected, because of a host of different reasons that we can't really get into now. And perhaps the more prominent subfamily of the consequentialist theories is called *utilitarianism*. And the idea is that what you should what you should do is take the best course of action. And the best course of action is that the mind as the one that has that causes the has the *greatest utility*. So, this utility, of course, whilst is a technical term of art for economists will use it to in their, you know, formal there is it's also open to debate when it comes to the philosophical positions about how it has to be spelled out, and of course, different utilitarians will tell you that there are different ways or with different ways to understand this notion of utility. Now, there are two, there are big sub families within utilitarianism. One is called *act utilitarianism*. And one is called *rule utilitarianism*.

In act utilitarianism, it's your single act that has to maximize utility, or as to have the best consequences where this best is then understood in different ways. But there are as you can imagine, a few a few problems already on the surface of this theory, namely, perhaps the most serious is that it's really, really *hard to foreseen the consequences of our actions in a global, complete way*. So when do you evaluate the consequences of our actions? Because an action that I took now has consequences in five minutes. So if I say, well, in five minutes, we're going to take a pause from the class. That's one point to evaluate the consequences. So is it good? If we take a pause or not? Does it produce more utility than if we didn't take a class and so on. But then suppose that during this break one of you, you know, falls in the well and dies? Well, I mean, it's of course unclear whether this effect is a direct result is a direct consequence or is a consequence of my action. But it comes a little bit later. Could I could I foreseen it? Did I intended so This notion of consequences of a single act are quite, quite hard to narrow down, it's a matter of distinction between the *moment of evaluation*, but also whether it's just the *intended consequences*, or the *real consequences*. So as you can see, there's a lot here to unpack lots of problems for this act consequentialism theory.

So, so another, perhaps more sophisticated consequentialist theory that people come up with, to offset this problem of evaluating the consequences of single acts is called *norm, or rule consequentialism*. Whereas, what you should do is those *actions that agree with the principle that has the best consequences*. So, you see, in a sense, it's this theory, it still is a consequentialist theory, because actions are judged, according to the consequences, but not the consequences of the single actions, the individual actions, but rather the consequences of the principle. So, we're still focusing on the consequences. But now, the consequences are the consequences of a principle. So, perhaps it's a bit more in line with deontology. But of course, not in the structure of the theory itself. Then there are like, funnier versions of consequentialist theories. For instance, *egoism* is a form of consequentialism. Because what should I do? Well, I should do those actions that have the best consequences for me. So of course, you judge what you should do with the consequences of actions, so it's a consequentialist theory. But the way you judge these actions when the best consequence sorry, the consequences, the best consequences are your own benefit. So let's say that perhaps this theory is very popular not as a theory but as a practice, but not very popular among ethicists.

Virtue Ethics

Under two theories, virtue ethics and particularism are perhaps a bit less common. Virtue ethics maintains that you should do what's virtuous. And what's virtuous is displayed by a virtuous action actor or agent. So you know, what? Suppose you're a soldier in battle. And what you have to tend and try to do is actually be a good soldier, a virtuous older. And then of course, you can discuss what constitutes what constitutes being a virtuous soldier, for instance, being courageous, or obeying orders or being compassionate, I don't know. But that's the general idea. Single out-of-character acts shouldn't be held against the actor, as they might not be following from their general character.

Particularism

The last first-order moral theory we're considering is *moral particularism*. Moral particularism holds that it is *not the case that there is a general principle of action*.

Thus, the theory doesn't hold that for any particular concrete situation, there's one good response to our general moral principles, not that you have to do your duty, not that you have to do the action with the best consequences. Not then you have to act in such a way as to embody virtue and so there there's nothing general in ethics but only the particular. So in a given situation, there are more morally relevant factors that are called *reasons*: some of them favor one side of the issue, some favor the opposite side. So, you should imagine a scale perhaps, and so, some of the morally relevant factors some of the reasons, points in in favor of a particular course of action. Some other factors that are also more relevant points, the words against or disfavor that course of action or perhaps favor a course of action that is incompatible with the first course of action. And, *in order to decide what to do, you have to go through a balancing or weighing of these reasons in favor and reasons against*. And then, of course, there are different versions of this theory too. But the idea is that, you end up with an overall or all things considered were pro total reason, either in favor a particular action or against a particular action that is more important or weighs more than other things, and that's what you should be doing. So, there are tons of problems with this account as well which perhaps, we might deal with in the future. But one thing that I wanted to point out is that *these reasons, the same reasons, can switch polarity in different situations*. So, the fact that one morally relevant factor is in favor of a course of conduct in one situation does not guarantee that the same or relevant factor is going to be in favor of that course of action in a different situation. So, this morally relevant factor can just have a different way weights or importance or could even switch polarity. And so, in a situation a to be in favor of action of action sign was in situation b, the same factor the same reason is against course of action five. So, in a sense more particularism is extremely contextual.

So, to recap, we've seen four different families of first order moral theories deontology, consequentialism, virtue ethics, and particularism. Deontology some tells you that you should, what you should do is, do what's right, what's your duty, there can be more than one duty or just one duty, one principle, according to which you have to regulate your action consequentialism instead evaluates the right action in accordance with the consequences that come out of it. And then of course, there are different ways you can evaluate these consequences with

reference to happiness or utility or something else. Virtue ethics does not consider like actions in a vague room, but rather the moral character of an agent and put the focus on virtues. Whereas particularism just tells you that there are no general moral principles, but what you should do there is from context to context. And the artist and it's the result of this kind of practical reasoning, this kind of balancing or waiting act between different, more irrelevant factors that can change the importance and polarity from one context to the next.

Metaethics

When it comes to ethics, we saw that we can distinguish three levels at which both ethical inquiry is usually performed *the linguistic level, the epistemological level, and the ontological level*. After World War Two, in English speaking philosophy, there's been a fair bit of what's called *ordinary language philosophy* movement or revolution. As a reaction to the criticism of metaphysics, done by the Neopositivists between the two world wars, who thought that the questions about metaphysics were totally nonsensical, and therefore not worth pursuing. Nowadays that attitude has largely been abandoned. Now metaphysical inquiry is one of many inquiries that is possible at the philosophical level. So historically speaking, metaethics as always been its own thing, considered separate from other kinds of ethical inquiries. Thus, the language level has been quite prominent. But what is the language level? The language level has to do with what intuitively or superficially we might identify with moral language. And that means the language we use to express moral claims, and to the language we use, and when we engage in communication about morality: e.g. when we say "giving to charities is good", or "keeping your promises is the right thing to do", and so on. This part of metaethics would then ask, okay, "what's the meaning of good?

" "What is the meaning of right?", "Does good imply right?", "Or perhaps vice versa, Right implies good". And so this is strictly so what's the semantics of moral language? No one could say, well, I mean, there's absolutely no grammatical level, especially when we use adjectives in good or bad, morally right. Rather than when we use *modal verbs*, like "you should know this", and "we ought to keep your promises", and so on, but could say, well, I don't really see an issue here. Because it

seems like *these sentences that you just uttered, grammatically speaking, display absolutely no difference to or from descriptive sentences*. So I might as well say, "Oh, the sun is shining today." And let's say about abortion, "abortion is wrong". So there's a subject: 'the sun', 'abortion'. There's a verb in this case, it's a copula 'is' in both cases, and then some sort of predicate: 'shiny', 'wrong'. So on the surface level, the grammatical level, it doesn't look like there's much difference between the sentences that we use to describe the world and the sentences that we use to convey moral messages.[1] Okay, well, but if that's the case, the thought goes, then a sentence like 'the sun is shining today' is *truth-functional*. What does it mean? Well, it means that under normal conditions, this sentence is either true or false. And in fact, can be verified or falsified by the world out there. In particular, the sentence 'the sun is shining today' is true just in case the sun is shining today, here and it's false if the sun is not shining today here that is stated that's a fact that is external to us. I mean, even if there are no humans, the sun today is still shining, unless you're a hardcore idealist. So, so if we took a moral sentence like abortion is wrong, as a descriptive piece of language, then we should be asking, Okay, what is it? So it's as to be either true or false. It needs to be able to be either verified, or falsified by something. Because, I mean, since there's no difference with descriptive language, and one expects the same truth conditions to apply also to this kind of world. So many people think that that's, that's dubious, at best, and a grave mistake, at worst, and why is that? Well, because many people think that there's nothing out there in the world like moral facts or in terms of the kinds of things that would verify or falsify sentences of the form of 'abortion is wrong'. So there's no fact of the matter that it's descriptive objective out there, that will make the sentence to recall us. Okay, but I mean what's the semantics of the sentences then? Because we do use them and in a pretty successful way and so there must be a key here, right? If, okay, so let's let's go along with the objectors and say: Okay, this sentence could not be true or false. And yet, I can use it in reasoning, that also has mixed premises. So I could say:

abortion is wrong.
If abortion is wrong, then getting your little sister to abort is wrong.
Therefore, getting your little sister to abort is wrong.

[1] Imagine how difficult it is for a machine to detect the difference, if it cannot rely on grammatical cues.

The conclusion is one of the premises is not moral because mean as a conditional you're only considering the truth of the matter. Or you're just not endorsing what's being said, you're just assuming it conditionally. And yet, you can say, well, the conclusion follows or doesn't follow. But it still makes sense to have this kind of reasoning. So there's a whole host of issues here that we can get into.

I would like to give you a bit of a panoramic view of the theories that do not hold that sentences like "abortion is wrong" are descriptive. And so they owe us an account of what they are, and what's their semantics.

It all started with an English philosopher, Stevenson, who was an *emotivist*: when I say the abortion is wrong, what I'm conveying is an emotion, something like 'Boo abortion!'. It means I don't like it. And when I say, keeping promises is good, when in fact I'm doing is saying, 'hooray for keeping promises!', that is: I like it.

Richard Mervyn Hare, another English philosopher, elucidated a bit this view, and also gave, in fact, an operationalization of it, and made it altogether much better and more refined. He first distinguished between the content, as it were, of a word and second its mood, its force to be either descriptive or expressive, as it were. And then, as a further way more contemporary refinement. This is called a *non-cognitivist position*. So, the position according to which moral sentences are not truth functional, and do not express descriptive knowledge as it were. Hence the non-cognitivist label.

The contemporary position that hold this view is called *Expressivism*. And Expressivism, well, it's one of the most important contemporary theories, is in fact best viewed *as a meta-semantical theory*, because it just gives an account of the meaning of sentences, in fact, both descriptive and normative. So they really do a good job in providing a unified view, and not just an ad hoc adaptation of non cognitivist position, just in moral sentences. So at the start a descriptive sentence expresses a state of mind. And the state of mind is propositional. And it's normally a belief and borrow sentence, or perhaps more generally in a normative sentence also expresses a state of mind. But this state of mind is not a belief. So it does not have a preposition ON nature. It could be desire and intent intention, a plan that varies a bit with the expressiveness you ask. But the idea is that they have descriptive sentences and normative sentences have structural similarities, which would also account for their surface level. identity as it were. But, so they both

express mental states, but one case the mental state is propositional and it's a belief. In the other case, the mental state is expressed by a normative sentence is not propositional. And it's not a belief. But of course, since we can do all sorts of operations or normative sentences exactly as we can do on descriptive sentences, and what I mean by this is that we can conjoin them, we can disjoint them, we can use them in unassertive contexts, like conditionals we can effect now, for descriptive sentences, that's not really a problem right, because since they express mental states that are propositional in nature, then you can just hypothesize that there are these operations on propositional mental content mental states and beliefs that in fact, mirror or structurally similar to the kinds of operations and connectors that you would have on prepositions. So, conjunction, disjunction, conditional and so on. But the problem is that we can do the same things with normative sentences as well. So, we can, as I said from join them to join them using in a certain context I conditionals. And now, the expressivists owe us a theory on how we can do this. And why is that well, because, if the content that they express is non propositional then they can't just avail themselves to propositional operations. Because the kind of mental state expressed by this preposition by the sentences and not the kinds of things that can be conjoined and disjoint and using conditionals with propositional connectives, so, they basically have to hypothesize and replicate all sorts of operations, something like conjunctions something like disjunction, something like a conditional, something like negation at the level of non propositional mental states, so, between plans or desires, and so on. So, this is not good in general for the expressiveness because that needs a lot of work. That gets fairly unintuitive. But there's another, perhaps bigger problem, which is the fact that it's not that you just have descriptive sentences and normative sentences, but *you also have mixes so you can have a normative premise, a descriptive premise and perhaps a normative conclusion*. Or you can even have a mixed conditional. And that's really a problem because what do these sentences express? It mixes between propositional mental states and non propositional mental states. This is particularly evident in a problem that it's called the Frege-Geach problem for expressivism, which is a bit of a hotbed a test case for expressivist theories.

Ontology

We have already seen that some semantical questions in metaethics depend on ontological matters. For instance, if one adopts a correspondentist theory of truth, then whether moral sentences are true or not depends on what facts exist out there, and in particular, if moral facts (or values) exist.

When it comes to the ontology of ethics, we only see the main theories: realism, anti-realism.

Realism, or objectivism, roughly holds that moral facts (or values) exist out there in the world, and they are objective. These may either be abstract properties, non reducible to other kinds of properties, or these may be somewhat emerging (supervening) from other, presumably physical, properties. For the non-reducibility version, one of the famous arguments is G.E.Moore "open question", that, among other things, purports to establish that good(ness) is not reducible to any other property. One last form of realism that I am mentioning is the one that considers moral properties dispositional, so not independent of anything else, but somewhat embedded in the fabric of reality (compare the property of being fragile).

Anti-realism, or relativism, holds that that moral facts or values do not exist out there in the world, or at any rate not in any objective form, although they may have a certain level of subjective existence, or perhaps of inter-subjective existence, as it is often the case for those who are called moral constructivists.

Epistemology

Epistemological issues in metaethics are mostly about how we attain (if we do) moral knowledge.

The biggest divide in the field is the one between *cognitivists*, who think that moral knowledge is exactly as all other sort of knowledge, i.e. mostly propositional and factive (e.g. this implies it is truth functional); whereas *non-cognitivists* think that moral knowledge is not propositional, but rather involving non-cognitive mental states: emotions, desires, or, in more advanced theories, plans.

There is another issue, having to do with how we justify moral beliefs. The most prominent position here is probably what is called *intuitionism*.

Moral Uncertainty

A very interesting, and very relevant issue also to the field of machine ethics, is the topic of moral uncertainty.

The problem is the following: *what should one do* (in terms of a first-order moral theory), *if one is not sure which first-order moral theory is correct?*

One prominent contemporary proposal,[2] roughly suggests that one should apply the methods of decision theory in the following way: one weighs their subjective credence on any single moral theory, and multiplies it for the value of that particular outcome according to that particular moral theory. A couple of issues with this proposal: first, the uncertainty here is subjective, thus it reflects one's convictions, biases, etc. rather than being objective. Thus, it is of limited use beyond the individual level. Second, while such a proposal might work in fairly homogeneous normative context (say within broadly consequentialist theories), it faces tough difficulties when applied to a more heterogeneous landscape, because it assumes intertheoretical comparability of the value/utility functions, something that cannot really be assumed for a deontological and consequentialist theory, for instance.

Applied Ethics

Applied ethics

Ethics of AI, Ethics in AI

Even before discussing ethical issues in and of AI, we can already ask a simple question: where does the ethics of AI fit in our tripartite scenario: normative ethics, metaethics, or applied ethics?

It turns out that a simple question does not necessarily have a simple answer.

[2] William MacAskill, Krister Bykvist, and Toby Ord, *Moral Uncertainty*, Oxford, Oxford University Press, 2020.

The simplest starting point is that some of it is applied ethics: how certain first-order moral questions apply to a specific domain, that of artificial intelligence. Thus, there will be a few problems on how to perform AI research, how to build certain algorithms, etc.

However, it seems that AI opens more fundamental questions for the field of ethics as well.
On one hand, there is an obvious application in *descriptive* ethics, as of course certain AI techniques can be used to gather and analyze ethical data and predict patterns.
But also at the normative level, there are new questions. First, can intelligent but non-human agents be moral patients? And can they be moral agents?
What structure ethics should have as to be understood or enforced by non-human intelligent agents?
Again, in case we reach an artificial general intelligence, or perhaps even the singularity, is it a good thing that humans loose control?
And if so, what kind of values will an AGI have? What values should it have?
All these questions are not questions of applied ethics.

Proper questions of machine ethics are about how to make intelligent machines at least behave in such way as to appear ethical.
I don't say "behave ethically", because at least according to some ethical theories (e.g. deontological), issues of motivations, intentions, etc are paramount to what constitutes behaving ethically, thus, for an AI to behave ethically with this standards, it would have to be a strong AI (in the sense of Searle).

One subdivision has machine ethics as a part of the ethics of technology. Other subfields of the ethics of technology would be robot ethics, which deals with how human beings design, construct and use robots; and computer ethics which is concerned with commercial behaviour involving computers and information (for example, data security, privacy issues).

Ethics in and of AI:

- Once we know ethical theories and the extent to which machines can be moral agents / patients (still open to me, really) the next question is:
- How can we build an **ethical artificial agent**? (at least on a purely theoretical level)
- We could build agents from different moral principles: Deontologist, particularist, consequentialist
    - Deontology in robotics: The "law of robotics" (arguably top-down, I think): there are certain principles that a moral agent must follow categorically.
    - Possible problems:
        - Misinterpret the rules
        - Lead to unforeseen scenarios
        - Problem already with ontology with humans: Going from a general principle to a concrete implementation is just really hard to do
    - Consequentialism: There are some elements in machine learning that resemble elements in consequentialist theory already: The reward function, evaluating possible outcomes, working with probability distributions and uncertainty…
    - Problems of an implementation:
        - To what extent are you in an environment that is closed to the extent that you have a chance of predicting the consequences of your action
        - The costs of "rerunning the programme"
        - How do you evaluate consequences
        - all the other problems with consequentialism.
    - Particularism: Bottom-up approach "Casuistry". You could train a network using ethical dilemmas, and then the system is supposed to solve new ones.
    - Problem:
        - Assumption in the training: It is questionable that there are "correct answers" to these dilemmas that the machine could train on. "Correct" according to who? Context might change a lot here as well

        - We would have to solve a lot of potentially problematic cases to do this and would still not know if we could trust the results
        - Example: *delphi.* When you reformulate your question in a way that does not change the semantics, it may heavily impact the response the engine gives.
    - Combined approaches: Now from the top-down: "MoralDM". Combining Deontology and Consequentialism into one model. Utilitarian reasoning applies until "sacred values" are concerned, at which point the system operates in a deontological mode and becomes less sensitive to the utility of actions and consequences
    - Other mixed approaches: The "Hybrid approach". Combining top-down (theory-driven reasoning) and bottom-up (reasoning shaped by evolution and learning) approaches to building moral agents.
- It seems to be: These different approaches are perhaps best seen on a plane. One axis: Bottom-up vs top-down (the approach of construction); the other axis: The kind of moral theory that stands behind. Yes, consequentialism and (particularly) particularism might mix better with the bottom-up approach, but even there, you could top-down the system by which value is assigned in a consequentialist theory or top-down the ways balancing is performed in a particularist theory…
- A societal problem: There is currently no agreement on what normative theory might be "the correct one". What theory are we then to implement in machines?
    - Different theories are not always compatible
    - Called "the problem of moral uncertainty": What should we do when we do not know what theory is correct
- Idea for a possible solution: The ethical knob.
    - Example with self-driving cars: Instead of hard-coding the answer to moral dilemmas like: "run over the grammie or drive off the cliff, killing the driver", the owner / driver decides what moral theory he wants his car to implement. Once chosen, this particular car will act in accordance with it in every dilemma.
- Another solution: Accept moral uncertainty. See above.

Autonomous Weapons Systems

Autonomous Weapon Systems already exist but are not necessarily lethal; moreover, they are not illegal according to the rules of war. Only some pose new ethical and legal challenges. Please refer to Nathan Wood's slides.

Existential Risk and Longtermism

An existential risk is an event that results in increasing the change of human extinction or some other unrecoverable global catastrophe.[3]

Thus, existential risk is not something that is limited to AI-related matters, such as the singularity or the control problem, but extend to ghastly events like nuclear disasters, deadly pandemics, climate change making the planet unhabitable, etc.

Should we mitigate and perhaps prevent existential risks?
The answer to this question requires we make explicit what are our values.
For one might say: who cares whether in 5000 years the planet is inhabitable to humans. People who are alive today will be long dead by then, and humans are a threat to other species and natural environments anyway. Or again: it is not worth it to change radically our way of living now, for scenarios far in the future.

On the other hand, someone might say: imagine the incredible number of lives that will be lived, and the enormous amount of pleasure and utility that will derive from them, if life on the planet continues as we know it. While these people are not alive yet, we have a duty to preserve and improve life on this planet for their sake, because there will be so many of them, that even discounting for the future will make it irrational and morally cruel to deprive them of any right to life etc.
This latter position is a rough phrasing of a philosophical position called *longtermism*.[4]

Longtermism is built on the assumption that future people matter, morally speaking, as much as people currently alive, and based on a reasonable assumption that there will be way more people in the future that have been alive so far (and *a fortiori*, in the present), we have a moral duty to take their well-being

[3] The literature on existential risk is by now pretty vast. I recommend Toby Ord, *The Precipice* (2021), as a starting point, and Nick Bostrom, *Superintelligence* (2014), when it comes to existential threats coming specifically from AI.
[4] Longtermism, under one name or another, has been around for ages. For a recent take, see William MacAskill, *What We Owe the Future,* 2022.

into account. Thus, in this sense, longtermism might argue for giving priority to improve the future rather than the present.

Even after such a brief characterization, there are several issues open.

1) First, there is the fact that long-termism seems inestricably linked to consequentialist ethical theories. This will make it harder to adopt for people endorsing other, less compatible with it, moral theories.

2) Second, there is a rather large theoretical and practical difficulty, which is how to measure future value.

3) Third, one can level more direct criticisms: deprioritization of current serious issues, reliance on vanishingly small probabilities, etc.

Current Uses of AI in the Law

Civil Law

Introduction

Today's class is on current uses of AI in the law, and in particular, outstanding legal issues that may arise when employing artificially intelligent systems in judicial environments.

Earlier, we made the distinction between *current AI techniques*, i.e., what's currently available or may reasonably be expected in the foreseeable future, versus what *might be available in the more distant future*. To reiterate, this is a distinction which will be made repeatedly throughout the course, as it calls into question different sets of problems and uses different terminology. For example, you can talk about *weak AI* versus *strong AI*, or *narrow AI* versus *general AI*.

The first elements of these pairs are what we are talking about. At a basic level, we are concerned with the issues generated by the use of current AI technologies, and how these technologies can be either designed or regulated to mitigate/eliminate the problems which they cause.

Today's class, very importantly, is about *civil matters*, rather than issues in the law in general. Thus, we are not looking at how AI may affect criminal law (this will be covered in future classes), but only how it relates to civil cases. As we have seen earlier in the course, the distinction between civil law and criminal law is, basically, the distinction between non-crimes and crimes. A non-crime is something like getting a parking ticket. This is in breach of some ordinance, but does not lead to prison time or being given the status of "criminal". A crime, on the other hand, is something more serious which may lead to sentencing to prison time. Our concern today is with the former only.

People interacting with the law

Before looking at the possibilities for regulating the use of AI in either the present or in the near future in the legal domain, let's first have a brief overview on how AI

systems are currently used by *people who interact with the law*. This includes practitioners of the law, like lawyers; administrators of the law, like judges, state officials, or police; and also users of the law, like everyday citizens and companies. For example, when you go to the train station and buy a train ticket, you are actually stipulating a contract. As such, you are making use of the law (of contract) in order to get somewhere on time. If something goes wrong during your train trip, then you can file a complaint and be reimbursed. The same holds for any number of basic transactions or interactions we have on a daily basis. And all these things, even the simple things of everyday life, are regulated by the law. So the law is not something that is out there and remote to us because we think, "When will I end up in court or go to a tribunal?" It is actually something that is in our life every day, impinging on many of our most elementary social interactions.

Litigation Discovery and document assembling

Very simply, if we start from practitioners of the law, i.e., lawyers or legal officers, companies which aid them, and so on, there is a kind of continuity between the use you can make of AI tools outside of the legal domain and those you can use within it. For example, AI tools can be used very simply to smartly access existing databases and have a technology-assisted review of the case at hand. This is called *litigation discovery*, and it is just the process of obtaining evidence for a lawsuit. Such evidence could be from archives and easily accessible, or it may be based on data collected by social media companies and bought or sold by third parties, making it more problematic as its procurement may infringe on people's right to privacy. Either way, as you all well know, when you have a lot of data these tools can speed up research a great amount. Thus, such AI tools allow practitioners to review a larger number of contracts than they could without them. You can also use AI tools as everyday legal aids for automatically putting contracts and other legal documents together, also called *document assembly*. In general, such use of AI-assisted legal research is nothing problematic.

Prediction of Legal Outcomes

Something more interesting (and problematic) is the use of AI technologies for the *"prediction" of legal outcomes* for some case at hand based on past cases which have been decided. More interesting, and perhaps more problematic still, is the use of AI tools by judges, the state administration, and the police. These tools can be used by

judges in helping with sentencing or bail decisions for criminal defendants. One case that has been very famous is an AI used in Wisconsin called COMPAS, that helps or helped with sentencing by taking into account certain variables, like living address, age, educational attainment, etc., to predict whether a particular person was more likely than not to commit a particular crime again. This would then be used to help in deciding how long the person should remain in jail. Interestingly, and this is a problem that will come up time and again, this COMPAS system is proprietary, that is, it is provided to the courts of Wisconsin by a private company. And the defendants, or the judges for that matter, as far as I know, could not access the algorithm that is at the core of the prediction system. This resultant *opaqueness* goes to the heart of some of the questions of *transparency* which surround AI technologies and their use. This opaqueness also calls into question plaintiffs' *rights to an explanation* for their sentence, something which we will deal with in the next few classes.

Use by government officials

When it comes instead to the use of AI tools by government officials, there is already much use in the *distribution of government benefits*. So, for instance, social housing or unemployment benefits have their distribution aided by AI tools in many instances.[5]

When it comes to policing, there is an AI-aided practice called *predictive policing* where machine learning technology is used to detect patterns from past crime data, in order to attempt to predict the location and time of future crime attempts. This may also be used to try to profile people who are more likely to commit certain crimes. Another aspect of AI-aided policing efforts which has generated much discussion, and that will also discuss later, is *facial recognition technology*. As you know, we just saw how widespread the use of surveillance technology in China is, but in general, this is also something that our police in Europe do as well. In fact, the recent EU AI Act proposes to regulate the use of facial recognition software, for instance, forbidding its use in real time, or only allowing it when there is evidence that a crime has been committed. Such regulation would make facial recognition technology usable only for identification purposes when a suspect for a crime has

[5] What are the pros and cons of such applications?

been identified, but would not permit its use for prevention or profiling of groups or individuals who are taken to be criminal, irrespective of whether or not there exist particular evidence indicating that. At any rate, these things can easily be abused and they are, unfortunately, barely regulated. It is also unclear to what extent you are surveilled or not when you are in a public space, or how long the data will (or may) be kept, and for what uses, and so on.

Smart contracts and expert systems

And then there is the use of AI tools by the *users/enablers of the law*, e.g., companies that help with complex compliance systems like tax filing. For instance, a large part of the financial market is made up of so-called "*smart contracts*". These are just securities contracts in the finance industry to see whether the trading contracts are expressed in computer understandable format and there are basically no humans in the loop for such operations. And when there is, for some reason, a glitch, the stock market may lose billions of dollars, and some of these trades cannot be reversed because they happen regularly, according to regulations. Thus, you cannot you just undo these trades. Due to issues of opacity it also may be the case that the errors leading to the glitch literally cannot be explained, even though it is very easy to see that something out of the ordinary must have happened.

Another more helpful and everyday example of an AI tool used by users of the law is what are called *legal self-help systems*. These are simple, expert systems that provide ordinary users with answers to basic legal questions. So suppose that you just got a parking ticket, and in trying to figure out what you must do, you access a chat box of the police department. Depending on whether you check a certain number of conditions, you can determine if you are permitted to appeal the fine, or if your fine may be discounted, or something else. This is a simple, almost mechanical example of an AI helping you to navigate your way to knowledge of the legal system, and that is where the AI comes into play; interactive features and the use of natural language allow a user to ask questions which are naturally answered, leading one toward some wished-for outcome, like learning if you can avoid paying a fine!

Issues with AI tech in the law

Let's now turn to a couple of issues, so that we may better understand why regulation is needed for current technologies being used in the law. The first is the clear potential for *bias in algorithmic decision-making*. If we go back to the example of COMPAS – the software that uses machine learning to predict the risk of criminals reoffending – this system creates its predictive model based upon past arrest records provided by the police. Now, let us first simply suppose that police activity in a certain area is itself biased. For instance, perhaps for some specific offense the police tend to arrest members of a certain ethnic minority at a disproportionately higher rate than non-minorities. This could be due to any number of issues and is not something which can be corrected by fiat. Moreover, such differences in who is punished may be due to the geographical organization of certain things or social goods, or because of more basic socio-economic issues. Thus, it is not even necessary that the police are acting on an unfounded or obviously objectionable bias, but even so, sadly, the skewed arrest records will embed bias at the heart of any algorithmic program which uses such data as an input. Simply put, any machine learning system that learns patterns from this data set is likely to, in turn, encode these biases. To see this, consider the following example:

Suppose you have a perfectly transparent AI system – and this is not necessarily the case with many techniques used in machine learning. In such a system, you see the algorithm itself, and you can ensure that it is not biased. So it is not that it says: if a crime has been committed and there is a white person, a black person, and an Asian person as suspects, then rank the black person as most probable, the white person as second most probable, and the Asian person as third most probable. Rather, the algorithm will look only at the previous arrest record for people of the same age, living in similar socio-demographic neighborhoods, in similar family situations, etc. Looking only at this "unbiased" data, the algorithm is then likely to say that the black person has a higher probability of being the perpetrator of the crime. This is because the data that were used for training the AI system were themselves based on biased human practices (that is, the biased policing efforts which led to skewed arrest numbers for some minority group). This is a crucial problem to bear in mind for any algorithmic tools used in the law: *human biases are very likely to lead to biased outputs, even if the algorithm itself is unbiased.*

The problem above is rooted in a slogan of computer science, "garbage in, garbage out". But this is not just about the epistemic quality of the data. It is really about problematic value patterns that are encoded in the data. So that is what bias is about. And even in this case, we assumed that the predictive system in question was transparent, so we can actually inspect the system and see how decisions are made. But is a very well-known issue that AI systems are often not interpretable and transparent, and how AI systems make their decisions is equally not interpretable or transparent. As a potential defendant, or as a citizen more generally, you do not want to be subjected to decisions that cannot be explained or motivated in our legal system. This is really one of the requirements for decisions to be valid in such a system; sentences from courts have to be motivated by the law and evidence of having broken it, decisions by administrative authorities have to be motivated by the purpose of those authorities and the evidence that justified them having acted as they did, and so on.

Another minor issue is the *deference to automated computerized decision-making* that comes as AI becomes more and more ingrained in government administration (something I am sure has affected all of you at some point as well). When you get some data, for instance, the speed on your car, and it does not align with your perception of how fast you are going, there is a tendency to say, "Oh, no, I must be wrong." After all, the speed of your car is measured by a precise tool, and it is natural to just defer to the tool in question. But suppose that it is a judge doing this, i.e., deferring to a "precise tool" which aids him in his legal decision-making, and he says, "Well, I know that I'm human, I'm limited, I might be biased, etc. If the algorithm suggests this, then there might be some merit and let's go with it." The judge then abdicates his judgment to that of an AI system, without even thinking about how or why the system came to that decision, and without even having thought that he had any responsibility to follow the AI's suggestion.

AI, Criminal Law and the Liability Gap

The topic of today is AI and criminal law. So we're going to review the essential elements of the criminal law, although in a fairly simplified form and we're going to at least introduce the topic of AI and criminal law.

In the past decade, there's been a lot of research, study, and speculation on AI and civil law; damages, tort law, and so on. And yet, there has been comparatively little attention paid to some of the significant problems that AI can generate in the criminal domain, that is the criminal law domain. Of course, there are already some applications of AI tools in the criminal law domain, for instance, help with sentencing, certain uses by the police, and so on. But there are also genuine issues. In particular, today, we will see two things; first, whether *there can be crimes committed by an artificial intelligence (or with an artificial intelligence)*, and second, *what to do about these if they do occur*. But let's start with a recap on what the criminal law is about.

Criminal law consists of rules about what is deemed minimally acceptable behavior in society, and punishments for breaking those rules. Criminal law is also about preserving certain goods that society deems beneficial to all, so this body of law relates not just to actions but also omissions, and even mere attempts to commit certain actions. For criminal law though, the mere presence of an action, omission, or attempt does suffice for being deemed worthy of punishment. Rather, these actions, omissions, or attempts must be accompanied by a specific mental state, called for instance *mens rea*. Depending on the country or jurisdiction you're in, these mental elements may be referred to in different ways, but the general idea is that for an action to be considered a crime, the action itself must be criminal in nature and you will have to have performed the action recklessly, or negligently, or with malice, etc. These further factors serve to ensure that those convicted of criminal acts were actually doing something wrong, and not just unlucky in some way. For example, in certain jurisdictions you are legally required to help people in need when there has been an automobile accident. This means that failing to stop and help constitutes a criminal act. However, what if it was dark and you couldn't see their vehicle? What if their car had veered off the road and into a small ravine

with tree cover, so it was hidden from view? In these cases, they are still just as in need of help, and you not helping them is still a criminal act. However, you don't see them, and may even be unlikely to be able to see them at all. Technically, you are committing a crime of omission by failing to stop, but you were not acting with recklessness, negligence, or malice. You simply could not see them and did not know they needed help. This makes it so you were not behaving in a criminal fashion, even though failure to stop is a criminal act; without the appropriate mental state, you are not liable to punishment.

And now we come to another essential element of criminal law, namely *responsibility*, because a precondition of punishment is that a perpetrator bears a fitting amount of responsibility for the crime. Responsibility depends on certain conditions that might be internal or external to the subject in question. As we have seen previously, it might crucially depend on the subject's freewill or cognitive capacities. What are these cognitive capacities? Obviously, there will be debates and nuances in the law as well as in philosophy, but a fairly general view is that these cognitive capacities are best understood as a matter of being able to respond to reasons. This may also be called "reason responsiveness". So an agent is responsible just in case he or she is capable of 1) recognizing and 2) responding to the reasons that bear on the situation. What kind of reasons must agents recognize and respond to? In general, both systemic reasons and practical reasons.

So far, all these conditions that we have seen ensure that criminal law is addressed only to humans (at least so far). This is because humans are the only entities which can sensibly be seen to possess one of the necessary mental states for an act to be deemed criminal. This in turn is due to the fact that humans alone possess a sufficiently developed degree of situational awareness and which are sufficiently capable of choosing courses of action in the knowledge of all that will or may follow from them. Moreover, as we have seen, punishment requires responsibility, and humans are (legally and philosophically speaking) the only entities which possess the level of reason responsiveness necessary for bearing criminal responsibility.

In discussing punishment, there is one question we have not yet covered, namely *why are people punished?* One of the central purposes of punishment is *deterrence*; i.e.,

to prevent the future commission of illegal actions by either the person being punished or others who witness the costs of illegal behavior. Punishment is generally thought to have a deterrence function, but in order for it to work, it must be addressed to people who can actually be deterred. People can generally be so influenced, which gives legal commands their normative "backbone", because agents know the corresponding sanctions which follow if certain rules are broken, and that knowledge leads to compliance. Non-human entities, on the other hand, so far cannot be properly deterred by punishments not directly temporally connected to misdeeds, and so all of the norms and sanctions of a criminal law system are ill-suited to guiding their behavior.

There is also *retribution*.
Most current legal systems employ a mix of the two.

Moving the discussion to non-human entities and the possibility of punishment, let's consider two kinds of cases exemplified by the following two examples.

> *Shopping Bot*: In the first example, we have an automated online shopping bot deployed to buy and sell things on the dark web. The bot buys a number of things, among which are illegal drugs.

> *AI Phisherman*: In the second example, an AI uses social media to gather sufficient information about an individual in order to create messages tailored to that individual. These messages are designed to look legitimate and speak to some personal interest of the targeted individual, and importantly, the messages have embedded links, which if clicked by the recipient, give some information to the phisher or provide the phisher with access to the targeted person's information or account. This information or access is then transferred to a real person in the real world, who uses that information for criminal activities like theft or fraud.

Here we have two types of cases: first, what might be thought of as *"normal" crimes*, i.e., things like buying illegal substances, which is criminal when committed by a human (in the required frame of mind), and second, what we might think of as

artificial intelligence crimes, i.e., criminal activities where an AI plays some essential role (even if the AI is not strictly speaking necessary for the criminal enterprise to function). In the first type of case, current law only considers humans as responsible parties, and therefore criminally liable. This means that if a bot were to buy illegal things entirely on its own, there would be no criminal activity, merely some sort of mistake which would need to be cleared up with the authorities. But the question then is whether there should be a crime in this case committed by the AI controlling the purchasing bot? And in the second case, where an AI greatly aids the criminal activity of theft or fraud? Buying illegal things, theft, and fraud are all real crimes already, but what about "artificial intelligent crimes"? Can an AI commit a crime, or must there always be a human bearer of responsibility? What if the human only has normal non-criminal interactions with the AI, but through some emergent behavior or properties the AI still acts in criminal ways? And if there are "artificial intelligence crimes", how are they performed, what is required for them?

In order for crimes to occur, there are two components that must be present. There is a factual component (the action/omission question about *what* happened), and there is a mental component (did the agent act or omit action in a reckless, negligent, or malicious manner). The factual component (called *actus reus*) concerns the physical aspects of the (potentially) criminal action. The mental component (called *mens rea*), on the other hand, focuses on whether or not the *actus reus* was done with a guilty mind. *Mens rea* is generally taken to be a necessary element for attribution of criminal responsibility (as opposed to merely civil or tort responsibility). With these distinctions in mind, we can say that AI systems can rather clearly act on the physical world, and therefore do things which would satisfy the attribution of *actus reus*. Whether or not the AI is embodied in a robotic device, or only in software, it will either way have various ways it can impact on the world. However, *mens rea* is not so clearly attributable, given that AIs do not have a "mind" at all, much less a "guilty mind". (Note that other non-human legal entities, like corporations, may be attributed *mens rea* in cases where the company has some policies that are obviously reckless or negligent. In these cases, the criminal liability may fall on the company, but the attribution of *mens rea* will often be located in the decisions of individuals acting on the company's behalf.) In order to find an explanation for why AIs cannot be attributed a *mens rea*, one might try to further analyze what goes into that

concept, arguing that *mens rea* has two components. The first component is *cognition* and the second is *volition*. Cognition is the agent's awareness of reality, gained through perception, conceptualization, knowledge, and so on. Volition, on the other hand, amounts to intention, willing to do something. Lagioia and Sartor argue that AI systems can in principle fulfill the cognition requirement since they interact with the world and must have an understanding of it (at least as it pertains to the AIs function).[6] The problem is with the volition requirement. *Can an AI have an "intention" to perform an act?*

Can an AI have an "intention" to perform an act?

There are different schools of thought on this, not least of all because there are different ways one can fully spell out how intention is to be best understood. Without committing ourselves to one view or another, we can at least say that there is some evidence of intent when an agent (artificial or otherwise) can foresee the probable outcomes of some action, judge how the probabilities compare against those for other actions, and then take an action in the expectation that *this* or *that* outcome will follow. If the agent is free to choose between multiple actions and opts for one, knowing what will likely follow, this provides evidence of intent, at least in certain cases (even if this does not tell the whole story for "intention"). If we take this interpretation, then under certain conditions *mens rea* can also be attributed to at least some AI systems, either because we can see that this evidence of intention is hardwired into the architecture of the system, or because we can presume that this is present judging by the AIs usual behavior. Importantly, *mens rea* is also not a certain scientifically available fact about humans, but must be judged based on similarly vague things as this (which is why legal systems have judges and juries). We cannot put people in a machine to see if they had a guilty mind; we use heuristics and reasoning, and ultimately presume that they intended to do something or not based on the circumstances and their external behavior.[7]

So far we have seen that it is possible for an AI system to engage in actions that would be criminal if a human were to perform them with certain mental states. But when an AI performs such actions, there is no human with the corresponding *mens*

[6] F. Lagioia and G. Sartor,

[7] Compare the whole debate about neuroscience and the law. For instance, cf. Faroldi 2014, Ch. XXX

rea necessary for attribution of criminal responsibility. This means that there is a *gap in responsibility*, and we as a society want to discourage or eliminate such gaps.

Two Strategies for Law on AI

AI presents a number of new possibilities, some of which could be problematic, and therefore we as a society must decide in principled ways how to regulate these emerging technologies. To illustrate this point, just imagine you go back 100 years and all of a sudden cars start to appear. Back then you did not need a driver's license to ride a horse around. You would just learn as you went, because speeds were not very high, and accidents were more rare. And even if you were in an accident, it was not likely to be terribly dangerous. So there wasn't a problem with people just riding horses without having proper training. Then suddenly there were cars, cars made out of metal which, if involved in an accident, could easily kill people. Streets also became busier, and while a horse can stop itself if there is an obstacle, a car will not, making it more important for drivers to be competent. So there was a sense of "new technology, new problem", and society reacted by establishing new regulations to constrain the possible bad effects of the technology. Most importantly, drivers had to obtain driving licenses, but also things like pedestrian crossings needed to be established, along with codes of traffic, and so on. In short, a host of regulations had to be created to constrain both the technology as well as its knock-on effects.

So what do we do when a new technology is already entering into our daily lives and is still not being regulated? There are two main types of approaches, *one more conservative* and the other *more creative*. The conservative approach is an *interpretative route*, where you just classify the new phenomena under existing law. So looking at the horse/car example, you would note that a car is a form of transportation like a horse, or maybe more like a bike. And since nothing special is required for someone to ride a horse on the public street, we could extrapolate this to cars and not require any regulation beyond the existing rules for riding horses. Then there's the more creative route, that is, the *legislative route*, where you in fact create new norms or new legal categories.

Returning to our topic of AI, the creative route – i.e., of creating new legislation – might take as a starting point the idea of asking *whether AIs should be granted more rights and responsibilities as their autonomy increases*. This might be modeled after how

the legal domain expanded and is expanding to give some limited rights to non-human animals. For example, in the course of the past half a century or so, more and more rights have been extended to non-human animals, such as that they have to be kept humanely, that they can only be killed in certain ways, that they have to be treated when they're sick, and so on. This is all good, but it is important to note that it wasn't the case until a only few decades ago. Thus, this is a clear case where we have extended more rights to some agents which previously did not have them (and maybe were taken to not be capable of having them). However, in this case we did not give the agents more responsibilities as well. Thus, if you have a horse and it kicks someone in the head and kills them, we do not hold the horse (legally) responsible, but rather you, the owner. More graphically, we do not imprison the horse for what it did, though we may put it down for reasons of public safety. The main point is just that acquiring more rights does not entail that one necessarily acquires more responsibility.

With regards to artificial agents, there is an open question about how we should view their rights and responsibilities as they acquire ever more autonomy. In the literature there are a variety of proposals. One proposal is that we should consider artificial agents *as if they were ambassadors*. So they have a certain sphere of autonomy, but they do not have their own will or their own intentions. Rather, they just execute to a better or worse degree what we tell them to do. This means that they have a kind of freedom concerning the ways they choose to execute their missions, but they do not have missions of their own, nor can they create such. Now, this preceding thought may have been true a couple of decades ago, but as techniques have emerged and technologies have improved, this is not true anymore, because sometimes artificial agents can deviate so much from the course of action that we instructed them with that, in effect, their actions become something totally new, unpredicted, and possibly even unforeseeable.

Another option that has been considered in the literature is to think of artificial agents as analogous to *slaves under classical Roman law*. Obviously, slaves were people, so they could act on their own volition, but they didn't have legal personhood themselves. Essentially, they were the property of their owners. However, if their owner so chose, a slave might be granted a stash of money which

they could use as they pleased, and this granted them a certain level of autonomy. So for instance, a slave could start a business or buy and sell things. And if the slave somehow caused damages they would have to repay those damages by using their "own" money, that is, that money they had been gifted by their owner. So where is the analogy to artificial agents here? Well, we consider these agents as not having legal personhood, so they cannot be responsible for what they do. However, in the event that they cause damages, the endowment they have been granted makes it as if they were already insured, so the person or company or whatever it is that has been damaged can just file an insurance claim, as it were, and be fully compensated.

These are interesting approaches, but they are *not adequate for a series of reasons*. First, many of the agents we're talking about are *not physically located*, so it would be difficult to decide what jurisdiction should be responsible for evaluating claims made against the agents. Moreover, because they are not uniquely physically located and are *also modular*, it becomes very hard to determine who is ultimately in charge of them. Who is the owner? At what point did which parts of the system do what?

When it comes to this topic, the current legal system offers a couple of options that we haven't explored yet, but that we have seen previously in the course when we were talking about responsibility. Recall that when we talked about responsibility, we discussed the normal cases of individual responsibility, but also the less common cases of things such as *group responsibilities* and, what we are interested in now, the so-called vicarious responsibility and also strict responsibility (or strict liability). Now, what do all these terms mean? Well, *vicarious responsibility* is the responsibility one has for the actions performed by someone else. The easiest case is a parent and their child; if the child breaks a window the parent (as the legal carer) is responsible to repay the damage to the owner of the window. This is the case even though the parent may have had no causal or other connection to the window breaking. There may also be more debatable forms of vicarious responsibilities, such as, for instance, the responsibilities the editor of a newspaper has for the opinions written in the newspaper by journalist employed to write for it. *Strict liability*, on the other hand, is the liability that people have for the consequences of

their actions, without any necessary connection to negligence or recklessness on the part of the liable party.

How do these two legal categories existing in current legislation help? The idea would be that since AI agents do not have a legal personhood, they would be treated similarly to children. So either the owner, programmer, or user of the AI agent would be responsible for any damages caused by the agent, even if the owner, programmer, or user could not foresee or did not intend those damages which were caused called. This more or less concludes our review of the first route, the interpretative route, where we are basically using categories and tools that already exist in our current legal system.

What about the second route, the more *creative route*, which relies on legislation to bring about the creation of new categories? The first approach here is to suppose that these *artificial agents do have some sort of legal personhood*, perhaps like corporations do. Now, one school of thought says no, that it is impossible to give legal personhood to these agents because there is a clear causal connection between the programmers or creators and the outcome. Even if not all the steps are clear or not all the steps are constructible, there is still a clear causal connection, and there is therefore no moment of genuine autonomy. (Also, the fundamental traits that you might want to associate to legal personhood are missing, something we will see later in the course.) However, though AI agents may not have all usual traits we take to be necessary for possessing legal personhood, their potential creativity and unpredictability makes it so that programmers and creators may be genuinely unable to account for all that may occur far down the line in the agent's functioning. This leads to a further reason for granting them legal personhood, namely that it is simply pragmatic to do so. Let us take for granted that we do not want to give up our AI agents because they simplify our lives so much. However, if we want to continue using them, it is sensible to have a legal basis for adjudications of responsibility and for the recovery of damages. By granting them legal personhood by fiat, certain thorny legal issues may be avoided, even though it may not be ontologically or philosophically valid to think of these things as true agents.

This is where we are so far with the legal debate, at least when it comes to civil matters, and it is important to remember that one of the reasons (if not the main reason) for regulating these agents is to *manage and mitigate the risks that they cause*. We have already seen some of these risks, such as risks of discrimination and unfairness in algorithmic systems of justice, and we will see more risks down the line as we move away from currently existing technology and toward future possibilities, eventually approaching strong AI. So, risks need to be prevented as much as possible, and if some sort of damage occurs, then there need to be clearly set ways for that to be remedied. In this vein, the legislative route that one of the most important regulatory bodies in the world, the EU, has decided to follow, is to *completely abandon the idea of responsibility based on legal personhood for artificially intelligent agents, and instead completely adopt a risk management approach*. In any case, the good thing is that a unified and harmonized framework for regulating artificial intelligence is being laid down. Moreover, the main benefit of the EU is the common market, and compliance is an important part of firms' access to that market. And since the EU is one of the richest largest markets in the world, and generally also the most strict in terms of regulation, firms from other large markets are incentivized to simply follow the EU standard as well, as this grants access to the EU common market and by extension virtually all other less strictly regulated ones. Thus, the EU regulations aim to take the legal tools they have and use them to foster economic initiative while mitigating consumer and society-level risks. This unified framework has been crystallized in the *proposed EU AI Act of April 2021*. Currently, it is still only a proposal, but if passed it would be a regulation applicable in all member states, and it remains to be seen if it will be ratified, and if so, how it will be implemented.

However, as potentially positive as the EU AI Act might be, it still completely ignores the previous debates on the potential legal personhood of AI agents and the different notions of responsibility that might be used for constraining AI agents, instead just moving towards a risk management system. This means that the risks need to be defined clearly, and what they do is they define different levels of risks, and for each level of risks there are certain requirements that need to be fulfilled by either the back end or the front end for AI agents, i.e., either the producers and programmers or the owners and users. So, if there is *risk which is simply unacceptable,*

then the activity is forbidden. For instance, real time facial recognition software is forbidden because the risk is deemed to be unacceptably high. For things which are *high risk*, but not unacceptably high, there are heavy obligatory requirements and a pre-conformity check. For *low risk and minimal risk* things there are just certain transparency requirements that are imposed on the creators and the users. Even before we begin to analyze the EU AI Act, there are so many open questions that one can ask, because we know what a notion of responsibility looks like and we know what the notion of personhood can imply because we've been employing these forever. The novel thing is applying these to artificial agents, some of which may sometimes act unpredictably. And so we could ask, if these agents are really so autonomous, what level of risk can be defined in order to cover every possibility? Are these requirements that are imposed on the programmers, creators, and users enough? How do you define AI in a functional way, in an ontological way? And also, if you just regulate what exists now, then what about the future? What about general AI? These are all questions that we'll deal with in the rest of the course.

Threats and Types of AI Crimes

Now that we have introduced the concept of artificial intelligence crimes and the liability gap that comes with those, we have begun to see ways of dealing with them, and whether there are solutions to some of the issues and questions which arise. To grapple more clearly with the problems that come with AI crimes, let's consider a pair of cases, both of which involve the death of a patient following some medical therapy delivered by medical robots.

- In the *first case*, the robots carry out the therapy according to existing and tested protocols, but the patient has an allergy the robots do not know about, and the allergic reaction to the therapy kills the patient. Similar to the first case, in the *second case*, the patient requires the same therapy and has the same allergy (i.e., will die if given the therapy), but in this instance the robot knows of the allergy. The robot performs the therapy anyway, "intentionally" killing the patient. Moreover, in the second case we don't know why the robot went ahead with the therapy knowing it would kill the patient. What do we do in such cases? What options and solutions have been proposed for regulating (or perhaps punishing) AI when things go so wrong, either by intent or by accident?

There are *two main strategies* that might be pursued. The first is to maintain that *the possibility of crimes committed by an AI does not entail that AI systems can themselves be subject to criminal law*. Crucially, the point is that AI systems are not to be treated as agents or addressees of the criminal law, because *only humans* – as users, creators, developers, or owners – *can be the subject of norms of criminal law*, or the recipient of appropriate punishments for violations of those norms. How these users, creators, developers, or owners are sanctioned, and to what extent, may be based on existing regulations about similar areas of legal regulation. The *second strategy* for dealing with AI crimes, and a strategy not yet implemented anywhere, is to consider the *AI systems themselves as subjects of criminal law*. Under this approach, AI systems could be treated similarly to how humans are treated for similar violations, with fines, restrictions to personal liberty, incarceration in correctional/re-educational

programs, etc. In their 2019 paper, Lagioia and Sartor[8] argue that the second approach, that AI systems might be subjected to criminal law and subsequent punishment, *requires that AI systems be considered as having personhood*. I disagree, but since we have already seen a similar debate about the legal personhood of AI system in civil law, let's examine the potential legal personhood of AI systems in the criminal domain.

As a first approximation, such a view of legal personhood would roughly correspond to the notion of legal capacity we have in the civil law system. And what would this legal personhood or capacity consist in? If we adopt one of the most famous definitions provided by one of the best-known philosophers of law of the 20th century, at least in continental Europe, Hans Kelsen, then *the person is understood as the totality of rights and obligation which have the behavior of a human being as its content* and thus form a unity which constitutes the center of all the system of legal regulations. To say that an entity is a person is thus just to say that this entity is or may become the bearer of rights and duties. So in general, for legal personhood, there is a thin and thick conception. On the *thin conception*, an entity is considered a person just in case there is at least one norm addressing the behavior of that entity, perhaps attributing to it rights or duties. On the *thick notion*, however, the behavior of that entity must be addressed by a set of norms corresponding to a large to the norms which are generally applicable to humans. Under this latter conception, the entity will therefore have similar entitlements and burdens as humans do.

Returning to the analysis of artificial intelligence crimes, we can ask two questions: *what are the fundamentally unique and plausible threats posed by artificial intelligence crimes, and what solutions are available for dealing with them?* The first question directs us to design preventive, mitigating, or addressing policies to handle these sorts of crimes, while the second points to the technological and legal solutions that have been suggested and analyzed so far in the literature.

In looking at this second question, I am following the chapters of King and co-authors in Floridi's 2021 book.

[8] Lagioia and Sartor,

In this work, the authors point out *four causes for concern with regards to AI systems: emergence, liability, monitoring, and psychology*. *Emergence* refers to the concern that while AI systems might be designed in certain ways, with reactions and behaviors we think will be predictable, it is possible that when "in the wild" these systems may act in more sophisticated ways than we expected (or even though we designed for). Such "emergent" behavior, since it was not expected, can be therefore lead to problematic states of affairs. *Liability* refers to what we have called the liability gap, and the concern is that artificial intelligence crimes could undermine existing liability models, thereby threatening the dissuasive and deterring power of the law. With regards to *monitoring*, there are three kinds of problem with monitoring artificial intelligence, which the authors call attribution, feasibility, and cross-system actions. Attribution is a problem because smart agents can act independently and autonomously, confusing attempts to trace accountability backward and determine who in fact committed the illegal act or omission. Cross-system actions are XXXX. Finally, *psychology* refers to the threat an AI affecting the mental state(s) of users in the general population, to the extent that the AI facilitates or causes some crime to occur. With these terms in place, let's see in what kinds or types of crimes an artificial intelligence can be a part.

The taxonomy is based on the empirical work already done by King and co-authors, who went through the criminal statutes of English and Welsh law and checked (and cross-checked) whether there was some law or precedence which might point to the existence of a crime of the sort which could be committed by artificial agents. Their finding was that there was certainly *a number of financial crimes that could be attributed to artificial agents*. For instance, forming cartels, price fixing, collusion, insider trading, market manipulation, and other such things can all perfectly well be executed by artificial agents, and in some of these there is perhaps an essential (or at least far more effective) role to be played by an AI. There can also exist drug-related crimes perpetrated by (or at least aided by) artificial agents; AIs have the potential to traffic, sell, buy, and possibly even possess illegal drugs. Moreover, there are some examples of *criminals using unmanned vehicles, which rely on AI planning and autonomous navigation techniques and technologies, to move drugs or other contraband*. This could be viewed as just another way to improve smuggling and distribution practices, but it still relies in its crucial functions on AI systems.

Moreover, since an unmanned vehicle can act independently of any operator, this can create responsibility gaps because criminals can avoid implications on the argument that they're not really involved in the transportation of the illegal drugs; since there's no link that can be ascertained between the criminal who sends drugs and the autonomous vehicle that transports them, especially if the software is sophisticated, it may even be impossible to establish criminal liability.

There is also the possibility for a *social media aspect of AI to be used illegally*, in the form of *bots fulfilling illicit functions*. For example, returning to the above example of drug trafficking, social bots might be used to locate and sell drugs to potential consumers. This use of bots would be illegal, but would make it more difficult to locate the human criminals ultimately responsible. There are also possibilities for malevolent actors to use AIs for crimes against persons, from mere harassment to things that could be considered a form of torture. For example, a malevolent actor can deploy a social bot as an instrument of direct and indirect harassment, generating fake content either under the name of the harassed individual, or targeted at them. Deep fakes, spam, and threats to reveal private information are also all possible to carry out with the aid of AI systems.

In addition to the crimes above, one can even imagine *scenarios having to do with torture*. The reason why AI could become essential in instances of torture is for the same reason AIs were used in the drug trafficking example above: *to create distance*. By adding a distance – moral, emotional, physical, and legal – between the agent ultimately responsible for the torture and the one being tortured makes it easier for that person who orders the torture to actually give that order. Moreover, the fact that AIs can't feel empathy or understand the pain of those they're torturing means that they might be more effective torturers than humans. And finally, by tasking the task of torture to an AI there is no person who physically goes to do that, and therefore there is no one who in fact committed an *actus reus*, or who can be brought to trial, or found guilty of having tortured someone, or punished for that. At least, that is the case under current definitions of criminal law.

A less obvious type of crime where artificial agents might impact on the situation are *sexual offenses*: rape (sex without consent), sexual assault, sexual touching

without consent, and sexual activity with a minor. Non-consent is in general thought to consist of two conditions. The victim must have not given consent, and the abuser must also lack a reasonable belief that consent was given. These conditions and their particularities can change, but that does not really matter for the purposes of our inquiry, nor does it matter how precisely these are formulated (although this is extremely important in real life). The potential issue with artificial agents is that advanced human-computer interactions can be used to promote sexual objectification, sexualized abuse/violence, and in a loose sense potentially simulate and hence heighten sexual desire for sexual offences. Artificial agents can therefore act as a sort of indirect promotion. For example, sex bots are supposed to have a humanoid form, the ability to move, some degree of artificial intelligence, the ability to interact with their environment, respond to stimuli and signals, and so on, and because of this it is entirely possible to simulate sexual assault or violence on a highly lifelike sex bot. Even more problematic still, some sex bots could be designed to simulate sexual offenses such as rape, or even child rape, and so on. So the use of anthropomorphic sex robots can have a sort of causal role in blunting people's natural aversions to sexual offences, or perhaps even changing those aversions to a desire to commit these crimes. On the other hand, there's an argument to be made that sex bots could serve a therapeutic purpose, given that they could be used to (re-)educate people with certain tendencies, thereby improving their lives and society at large. These discussions of sexual offenses represents an area where there are still many open questions about artificial agents.

The last set of crimes the authors consider and are *crimes involving theft, fraud, forgery, and similar acts*. The idea in these is that an AI could be used to gather personal data, and then use that data (and other AI methods) to fabricate an identity that could convince banking authorities to make a transaction. Such data could be gathered using social media bots that target users at large scale and low cost. Using such bots, one can build a rapport with the victim and manipulate them and their behavior, exploiting the established relationship to either obtain information or access a computer (recall the automatic phishing we saw before). There is potentially no real limit to this either, as facial recognition and voice recognition (and reproduction) can create increasingly deceptive ruses. Thus, *artificial agents create new and unique ways to commit everyday crimes of a variety of natures.*

Can an Artificial Agent Be Responsible? Problems and Models

So far we have seen some profiles or types of potential criminal activity either performed with the aid of AI tools, or done by an AI itself. Now we will examine possible solutions. Following King et al., we have identified our four causes of concern: emergence, liability, monitoring, and psychology. Using this, we could identify solutions to each of these concerns in turn, but we could also do something different, and instead try to take a top-down approach. On this approach, we ask what ought to be done if we want to prevent harm being caused in the first place, whether that harm is brought about innocently with AI tools or is the result of intentional or reckless action.

Given that we are talking about the criminal domain, our core question is whether AI agents can be responsible (either in a forward- or backward-looking sense)[9] for criminal actions, where this notion of responsibility can be understood as an agent's ability to be subject to blame or consequences (i.e., re-education). So first, following Lagioia and Sartor,[10] we distinguish two different things: how to make a system avoid innocently causing harm, and how to make a system avoid intentionally or recklessly causing harm.

In the first case, to *avoid the innocent imposition of harm*, there are at least two things we could do. *First, we could restrict or limit systems in what they can do*, either in their functional capacities, in their spheres of action, in their autonomy, or in their deployment. *Second*, we could *endow artificial agents with superior cognitive capacities so that they can figure out the unintended effects of their actions*. This second approach sounds a bit utopian, given that it assumes (or appears to assume) that an agent has superhuman abilities. However, there is a difference between having capacities greater than those of humans and having capacities equal to those of God; there is a middle ground that might be taken. An additional objection is that this second approach seems to implicitly assume that the universe is completely deterministic

[9] Responsibility can be backward looking, i.e. be focused on what has happened in the past in a retributive spirit, or forward-looking, i.e. taking into account what happened in the past in a consequentialist spirit, i.e. to avoid that it happens again. On this, cf. Faroldi 2014.

[10] Lagioia and Sartor 2019.

and every course of action can be predicted. But this doesn't really take into account all the different variables that are at play here, because there are also other agents, and there is such an enormous amount of variables that perhaps you really do need to be omnipotent to figure out all the unintended effects of your actions. All that being said though, it is at least plausible that advanced AI agents can figure out more effects than a human might be able to, even if it too faces limitations. This increased ability to predict outcomes may well increase the burden of responsibility for advanced AI agents, but this is also what happens with human agents; those capable of figuring out the effects of their actions are expected to utilize such abilities or be found negligent or reckless in their behavior. Thus, as cognitive resources or knowledge increases, the burden of responsibility will increase as well, and this would also be the same for AI agents. So to recap, we might prevent a system from unintentionally causing harm by either limiting the system (so it lacks the capacity to harm) or augmenting the system (so it can locate potential sources of harm and avoid those itself).

The second case focuses on *how to avoid, mitigate, or prevent intentional or reckless harm*. One of the first things that can be important for this aim is to ensure that systems do not adopt malicious attitudes, meaning that they do not actively seek to harm individuals. Ideally, such limitations are also imposed without also limiting the system's capacities, autonomy, sphere of actions, and so on, as this would make such systems less capable of carrying out the various functions users want them to undertake. But how is this to be accomplished, to ensure systems are non-malicious while still maximally capable? Lagioia and Sartor propose *two different strategies*. The first is to *provide appropriate disincentives for malicious action to either developers or users*, or possibly even the autonomous systems themselves. This is routinely done in training situations, where an operator, developer, or user specifies certain behaviors which they wish the system to avoid, and then imposes negative consequences when that behavior manifests. Such pairings of negative consequences with unwanted behaviors need not be limited to training environments, but can also be employed after a system has been deployed, as a sort of punishment system (fitting with the way criminal law works more generally). The second strategy proposed is not punishment-based, but instead relies on *endowing the system itself with a normative architecture*. What this means is that,

essentially, we build systems which can respond to norms in a genuine way, and not just as regularities or behaviors that are learned by rote or probabilistic success metrics. Thus, we aim to have responses to values and norms which are rooted in the normative, rather than statistical significance, of those values and norms. Later we will look more closely into what it takes to create a "normative agent", but for now let us see how these possible solutions might address the four concerns of emergence, liability, monitoring, and psychology.

Clearly, some strategies will be more apt to addressing some concerns rather than others, so let us go through them each in turn. With regards to emergent behaviors, a simple solution might be to limit the agent's autonomy or its deployment, as this would prevent it from having to navigate novel normative terrain which could lead to emergent or unpredictable behavior. Another approach would be to build normative agents, requiring developers to ensure that artificial agents have runtime legal compliance layers firmly in place before such systems are deployed. This would amount to creating legally binding rules and imposing constraints on developers regarding the runtime behavior of the agents they are creating. Thus by use of certain predefined legal rules, this technical solution proposes to regiment compliance by making non-compliance practically impossible in normal settings. However, here an immediate objection can be raised regarding the distinction between regulation and hard-coded behavior. Regulation leaves room for deviation from a norm, even if it sets consequences and sanctions for such deviation. Thus, compliance is still very much a voluntary affair; one can comply in order to avoid a sanction or choose not to comply in the knowledge that this imposes a large risk of being sanctioned. However, compliance is still up to the individual, so to speak. Yet hard-coding behaviors into or out of artificial agents also has risks, given that sometimes breaches of a norm or rule can be permissible or even obligatory. Consider a traffic light, for example. When it is red you are supposed to stop. But not necessarily in all cases. If a policeman is in the intersection and waives you onward, his direction overrides the normative force of the light. Or if you are racing someone to the hospital who may die if not brought quickly enough, then necessity overrides traffic concerns (within reason). So hard-coding limits may work as a broad-stroke approach, but it inherently creates other problems down the line. The problem is well described by Hildebrand, who makes clear that,

"While computer code generates a kind of normativity similar to law, it lacks, precisely because it's not law, the possibility of contesting its application in a court of law. This is a major deficit in the relationship between law, technology, and democracy."[11]

Supposing we do leave artificial agents with the ability to contravene the law, but we instill or create in them a certain fear of doing so, we still have to ask *whether AI agents can really be criminally responsible*. According to one theory of what it takes for agents in general to be responsible, the core requirement is a *sufficient level of reason-responsiveness*, i.e., a sufficient understanding of the epistemic and practical reasons at stake and a sufficient capacity to react to those reasons. Let us split this into sub-discussions.

We know that criminal law aims to discourage unwanted behavior through the threat of sanctions which negatively affect the interests of the agents in question. This is what we might call deterrence or criminal deterrence. Does this or can this also apply to AI systems? One preliminary question which must be asked is whether one can tell whether an AI system is or can be aware of its interests, or perhaps the interests of its owner or user and the ways that deterrence might impact on these interests. An immediate problem that arises is a known issue in the philosophy of law, namely the *distinction between sanctions and prices*. In order for something to be properly deterrent, it must not just be a price tag attached to certain behaviors. As an example, if traffic violations only ever result in fines, then these will rarely be deterrent for ultra-wealthy road users. Rather, for justice and deterrence to work, the punishment must both fit the crime and also fit the perpetrator. For example, a very poor road user may be deterred by a 5€ fine, whereas a very wealthy one may need a 5,000€ fine, or a 5,000,000€ fine, or may only be deterred by imprisonment or suspension of driving privileges. AI systems processing reasons of instrumental rationality will be directed only to maximize their owner's or user's utility, and this may make them view all sanctions as mere prices, calculating the costs and benefits

[11] Mireille Hildebrandt (2008, 175)

of (non-)compliance, and only complying when it is in the best interests of the owner/user. Thus, *AI agents may be incapable of being responsive to moral or other normative reasons*.

At this point, we need to distinguish between two things: 1) the mere ability to recognize norms in force in a society and the possibility of being subjected to punishment in the event of violations, and 2) a disposition to comply with norms purely on the basis of their conventional legitimacy or moral merit. Lagioia and Sartor conclude that *AI systems need to be responsive to both moral and legal reasons, that is, they must possess normative architecture*. This architecture is moreover understood as *the capacity to take values and norms into account*, and not be guided only by sanctions. However, if an AI agent needs to be reason-responsive for the purposes of criminal law, then there are *three capacities that are relevant here*. The first capacity is *the system's ability to gain or have awareness of its conduct and of the resulting effects of that*. The second capacity is *the system's ability to identify and understand the norms that apply to it, the sanctions that correspond to non-compliance, and the impact sanction have on the AI agent's interests*. And the third capacity is *the system's "moral motivation" to comply, which may be designed for or trained in a variety of ways*.

With all of this in place, let's suppose that we can create some sort of criminal liability for artificial agents. This brings us to the second concern of our initial four, namely liability. In the literature there are at least *four types of liability* that are discussed (following the early contribution of Hallevy[12]): i) direct liability, ii) perpetration by another (which I call in my work carer's responsibility), iii) command responsibility, and iv) actual probable consequence responsibility. The first model, *direct liability*, ascribes the fundamental elements of responsibility to an artificial agent and is what we might think of as the "full deal". This is a total shift away from the anthropocentric view of artificial agents as tools, where we now consider them as potentially equal decision-makers. Needless to say, this

[12] Hallevy 2008

approach is not what we see in current legislation; artificial agents do not currently have a separate legal personality or agency and therefore cannot be held legally liable in their own capacity. Naturally, this also means that they cannot contest verdicts, as they are not taken to meet the mental or psychological requirements necessary for having the legal standing to do such things. Interestingly, one related issue is that artificial agents are taken to lack a *mens rea*, but this is not always a necessary component for criminal liability, as we have seen with corporate crimes (where *mens rea* also obviously cannot be present or at least not demonstrated, but where criminal liability can obtain). However, there is a distinction between corporations and artificial agents, namely that in current legal frameworks the latter lack legal personality. An interesting point about the desirability of changing this status quo is that by holding artificial agents solely liable, we may create a situation where the human agents behind artificial agents (e.g., producers, owners, users, etc.) are stripped of any sense of responsibility for the AI agents' actions. This in turn can have undesirable effects on criminal law by weakening the dissuasive power it. To address this, one proposal of Floridi is indeed that the burden of liabilities be shifted onto humans and corporate or other legal agents whose decisions have made a negative difference on the actions of the various engineers, users, vendors, and so forth. So if a design is poor and the outcome is faulty, then all the human agents involved are responsible. The second model that Hallevy discusses, i.e., *the perpetration by another* model, uses an infant as the standard of *mens rea* which is to be attached to artificial agents. As such, artificial agents are treated as instruments for crime, but not as proper agents, and the real perpetrator is taken to be the party (or parties) who actually intended for the crime to be committed or created the possibility for the crime to occur. On this model there will be three candidates who might be brought before a criminal court in the event of an AI crime occurring: the programmers, manufacturers, and/or users. This places the burden of compliance at a variety of levels, as designers and programmers have incentives to make sure that artificial agents will not comply with illegal orders, manufacturers have incentive to not make artificial agents which they know might comply with illegal orders, and users have incentive to not give illegal orders (on the off chance that both the programmers and manufacturers failed in their responsibilities). If this model functions correctly, then a user of an artificial system must clearly intend harm because they have to

override the default position of the system in order for it to be physically possible for it to harm. This thereby also makes the model reduce ambiguity in court proceedings. The third model is *command responsibility*, which uses knowledge as the standard of *mens rea*. For example, in a military environment, this model would assign responsibility to any officer who knew about and did nothing to stop crimes being committed by soldiers under his command. This model is suited for use in contexts where there is a clear chain of command. However, it is unfortunately not readily applicable due to different complexities in programming, complex relationships between robots and humans, and other interfering factors. The last model that Hallevy considers is the *natural probable consequence liability* model. This model uses negligence or recklessness as a standard of *mens rea*, and applies to cases where a developer or user of an artificial agent did not intend some harm, but where that harm is a natural and probable consequence of their conduct. In such instances, they are then treated as having acted recklessly or negligently, exposing others to risk. Of course, there are issues with this model as well, as it is not always clear which programmer was responsible for which line of code, and it can be difficult to determine even which bits of code or hardware resulted in the problematic behavior.

Algorithm transparency and deep learning.

Logical notes in the margin of the European Commission's proposed Artificial Intelligence Act regulation and a Supreme Court order.[13]

Man out of laziness desires pure mechanism or pure magic.
Novalis
Only ignorance of the future makes the present bearable for us
Tiresias, in *Death of the Pythia*, Dürrenmatt

Summary

Introduction

The use of artificial intelligence is spreading to make some decisions that affect humans, whether in an instrumental or supportive way (e.g., for medical diagnoses),[14] or sometimes, or potentially, in an essential or autonomous way (e.g., in determining whether to grant a bank loan, or in determining recidivism profile).

The response of society, and with it the law, was not long in coming.

[13] Federico L.G. Faroldi, Senior Researcher, Research Foundation, Flanders (FWO) and Center for Logic and Philosophy of Science, Ghent University, Belgium federico.faroldi@ugent.be
[14] See, for example, GERKE et al (2020).

In this article I focus on two essential points: the notion of transparency and the logic of algorithmic decision-making, both as they emerge in some recent regulations and ordinances and as they may (or may not) be implemented technically.[15]

In particular, I will consider the GDPR, the European Commission's proposed Artificial Intelligence Act of 2021, and a recent Supreme Court order.
The GDPR prohibits automated decisions that have consequences for individuals, and establishes a right to meaningful information about the logic adopted in these procedures.
The European Commission's proposed Artificial Intelligence Act 2021 proposes to establish transparency requirements for algorithms, and provides an explication.
A recent Supreme Court order emphasizes access to the elements an algorithm uses in a decision and the enforcement scheme in which the algorithm in question expresses itself.

What is meant by the transparency of an algorithm? Is it always possible, at least potentially, to make an algorithm transparent?

This paper proposes a logical analysis of these issues. Specifically, in Section 2 I briefly introduce some of the terms of the issue, and in particular those of 'algorithm' and 'transparency' ('opacity'). In Section 3 I review the transparency requirements found in the GDPR, in the European Commission's April 2021 proposal for a regulation on artificial intelligence, and in a Supreme Court order filed in May 2021. Finally, in section 4, I start from the idea of making explicit the transparency of the algorithm with the logic used by the algorithm itself and an I propose an automatic technique for transforming an algorithm that is not at all or not very transparent (a numerical AI system) into a symbolic (and therefore more transparent) one,

[15] Both the judgments and the commentary literature are endless, and it is not possible to keep track of them in their entirety. Some recent contributions with regard to the Italian legal context are ALPA 2020 and DORIGO 2020. For recent works of comprehensive assessment of the issues involved, see RUSSELL 2019 and CHRISTIAN 2020.

moreover already designed to show plastically the justifications and reasons for the inferences performed.

The transparency of the algorithm

While it is very difficult to agree on a definition of artificial intelligence, there is a common distinction between reasoning systems and learning systems.

Reasoning systems are based on symbolic representations of source data and inference rules. They are logical systems, requiring initial organization of knowledge.

Learning systems (machine learning, deep learning, decision trees) are also used in the absence of symbolic organization of source data and problems. They are used, for example, for language learning, vision, behavior learning and prediction, and are largely based on algorithms and numerical techniques.[16]

In the first case, which is based on symbolic representations and logical inferences, the problem of both transparency and the logic of operation does not arise;[17] it is the second case, however, that is problematic in terms of both transparency and the logic of operation.

What is an algorithm?

Briefly, an algorithm is a precise, finite procedure that specifies how to obtain output data from input data. If there is a solution, it must be reached in a finite number of steps, and it must be clear whether the result has been obtained and therefore the procedure is finished.

An algorithm requires no creativity and can be executed automatically.

An algorithm can be efficient, if it finishes within a reasonable time, and reliable, if it produces errors in a minimum percentage.[18]

[16] RUSSELL and NORVIG (2020).

[17] At least as far as automatic processing is concerned; as far as creating the system is concerned, the issue is debated.

[18] There is obviously some debate about what precisely constitutes an algorithm. The features of this brief presentation follow NUMERICAL 2020.

But is it always possible, potentially, to make an algorithm transparent?
What is meant by the transparency of an algorithm? It is useful in this context to address this question by investigating the term antonym, that is, by asking what the opacity of an algorithm is.

At least *three* senses of opacity of an algorithm can be distinguished.
There is *intentional* opacity when the algorithm is not voluntarily made public in order to maintain a competitive or economic advantage. This type of opacity could be "solved" by simply publishing the algorithms in question--except for subsequent problems with competitiveness, etc.

There is cognitive opacity when the algorithm, while transparent, simply cannot be interpreted by most users because of their ignorance. In this case moving to an open system the opacity is not resolved *ipso facto*: an intermediate category of interpreters, explicators, or massive investment on educating the population is needed.

There is inherent or *essential* opacity when it is the algorithms themselves, net of programmers and publication, that make decisions in a way that cannot be explained or interpreted by a human, net of cognitive ability. This is the case with some machine learning techniques, and deep neural networks.[19]

For the latter sense of opacity in the computer science literature, the notion of 'intepretability' is often referred to.[20]

---

[19] For such a reconstruction, see BURRELL (2016). There are also algorithms that are totally transparent because of their architecture, which makes them self-explanatory, such as linear regression, decision trees or rule-based systems.

[20] ATARC, a U.S. nonprofit that facilitates collaboration between the federal government and industry, proposes a *transparency assessment* for an algorithm that takes into account the following factors, each expressed as a rating from 1 (not at all transparent) to 5 (totally transparent):

1. Algorithm Explainability

LIPTON (2016) gives a systematic account of the relevant literature, and argues that the interpretability of models falls into two broad categories: that of transparency *stricto sensu* and that of post-hoc explanations. In the first, what matters is to understand the mechanism by which the model works. In the second, what matters is extracting information from the models to clarify what exactly they have learned. *Transparency stricto sensu* can be divided into three levels, each referring to a different part of a model:

1. *Simulability*: a human can take inputs and go through model calculations in a reasonable time;
2. *Decomposability*: each part of the model (input, parameters, calculation) has an intuitive explanation;
3. *Algorithmic transparency*: has to do with the convergence or behavior of the algorithm.[21]

Lipton goes on to give some examples of post-hoc explanations:

1. *Textual explanations*: Similar to how humans verbally justify decisions.
2. *Visualization*: Qualitative determination of how the model is learned.
3. *Local explanations*: Displays what a network depends on locally, which is useful when global explanations are difficult to produce succinctly (saliency maps)
4. *Explanation by example*: Similar to how humans justify actions by analogy (e.g., word representations learned after word2vec training).

2. Identification of Data Sources used for Training
3. Methods used for Data Selection
4. Identification of Data Set Bias and Methods used for Reduction
5. Method and means by which model will be versioned

The most obscure point is the first. The definition provided of *explainability* is as follows:

"property of an AI system that important factors influencing the prediction decision can be expressed in a way that humans would understand (ATARC Model Transparency Assessment, 2020: https://atarc.org/wp-content/uploads/2020/10/atarc_model_transparency_assessment-FINAL-092020-2.docx)."

[21] This tripartition is from LIPTON 2016.

KRISHNAN (2020), however, argues more radically that the concept of 'intepretability,' is vague: thus it is almost impossible to know whether a given technical solution is acceptable or not; and in any case often that of 'interpretability' is a means to achieve other ends (such as non-discrimination or justification). We should therefore not force algorithms to be interpretable, but focus the discussion on the real ends we want to achieve.

Given the multiple realizability and vagueness, it is unclear how these notions of transparency found in the technical literature can harmonize with what is or will be required in artificial intelligence legislation. I will introduce this second aspect in the next section.

Transparency in artificial intelligence law

Transparency and explanation requirements are assumed in some regulatory sources that are already in force or will be in some form soon, as well as in case law production. I review some particularly significant junctures: the GDPR, the European Commission's proposed Artificial Intelligence Act 2021 regulation, and a recent Supreme Court order.[22]

It is well known that the first paragraph of Article 22 of the GDPR prohibits automatic decisions that have significant effects on individuals. More precisely:

> The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

There are significant exceptions established by Paragraph 2. But for our purposes more interesting is Paragraph 3, which establishes safeguards:

[22] While the European Commission's proposed regulation and the Supreme Court order in question are too close to the writing of this article for there to already be relevant scholarly literature, the output on the GDPR is already extensive.

> the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

According to some of the literature,[23] there would be a right to explanation [*right to explanation*] codified by the GDPR, according to others there would be at most a right to be informed.[24]

The most interesting feature, however, we can find outside Article 22, more specifically in Articles 13, 14, 15, which establish (among other things), at least in cases of automatic decision (regulated by Article 22), the right to have "meaningful information about the logic involved."

In the same direction is the recent proposal from the European Commission,[25] which thus generally summarizes the transparency requirements that are required for certain applications of artificial intelligence:

> *TRANSPARENCY OBLIGATIONS FOR CERTAIN AI SYSTEMS (TITLE IV)*
> Title IV concerns certain AI systems to take account of the specific risks of manipulation they pose. Transparency obligations will apply for systems that (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content ('deep fakes'). When persons interact

[23] GOODMAN and FLAXMAN (2016).

[24] WACHTER et al (2017). More generally for example against the "*explicability*" [*explicability*] of AI tools in medicine, see for example BABIC et al (2021).

[25] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS
COM/2021/206 final

> with an AI system or their emotions or characteristics are recognized through automated means, people must be informed of that circumstance. If an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, there should be an obligation to disclose that the content is generated through automated means, subject to exceptions for legitimate purposes (law enforcement, freedom of expression). This allows persons to make informed choices or step back from a given situation.

More specifically, here is how the transparency requirements are specified in the preamble:

> (47)To address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate.

It seems that a post hoc explanation is required, rather than the possibility that the algorithm itself is transparent in its inner workings.

~~This seems to be in tension with a recent Supreme Court order.~~
~~Indeed, Order No. 14381/21, filed on May 25, 2021, reads:~~

> ~~The lack of transparency of the algorithm employed for the specific purpose was not well disavowed by the contested judgment, which simply deemed the doubts regarding the automated calculation system for the definition of the reputational rating as not decisive, on the grounds that the validity of the formula would concern 'the evaluative moment of the procedure,' against which it would instead be up to the market 'to establish the effectiveness and goodness of the result or service~~

> provided by the platform.' This rationale cannot be legally shared, as the problem was not (and is not) confined to the perimeter of the market's response - a metaphorical synthesis to indicate the place and time when trade at the most various levels is carried out - with respect to the preparation of the ratings attributed to different operators. Instead, the problem, for the lawfulness of the processing, was (and is) constituted by the validity - precisely - of the consent assumed to have been given at the time of accession. And it cannot logically be asserted that adherence to a platform by members also includes the acceptance of an automated system, which makes use of an algorithm, for the objective evaluation of personal data, where *the executive scheme in which the algorithm is expressed and the elements considered for this purpose are not made recognizable* (italics mine).

The two elements of general interest are as follows: the requirement to make recognizable the executive scheme in which the algorithm is expressed, and the elements considered for this purpose.

With regard to the *first* point of the ordinance that is considered in this context, namely, the request to make recognizable the executive scheme in which the algorithm is expressed, we can identify it with the internal logic of the algorithm.

Regarding the *second* point of the ordinance, that is, that the elements considered by the algorithm be made learnability, it should be considered that the problem of *learnability*, that is, the selection of elements that the algorithm uses to make a prediction or make a decision cannot in general be considered decidable, for logico-mathematical reasons.

BEN-DAVID et al (2019) showed that the problem of *learnability* is equivalent to the continuum hypothesis, and therefore undecidable. In other words, *learnability* cannot be characterized in completely general terms.[26]

[26] Part of the reason this result, which is otherwise very elegant mathematically, is due to the fact that *learnability* is, in the literature, specified in terms of the existence of certain functions, rather

Two observations regarding the second point.
First, while it is reasonable for the requirement that the elements considered by the algorithm be made recognizable to coincide with *learnability*, this identification cannot be taken for granted in the absence of greater clarity on the part of the legislature, from a *de iure condendo* perspective, or the court.

Second, this mathematical result of Ben-David and colleagues does not detract from the fact that in some cases the selection of elements considered is not well isolable and knowable-even if for contingent reasons, so to speak (simplicity of the objective, etc.).

The Transparent Logic of the Algorithm

The first point of the recent Supreme Court order emphasizes the logic of the algorithm. In the latter section I show how the notion of transparency can be declined as information about the logic used by the algorithm. Specifically in my proposal, the logic used by the algorithm is not only readable by a human being, being symbolic logic, but is transparent to itself, that is, it makes explicit in the object language the reasons for various inferences.

This proposal pursues the idea of having an automatic technique for transforming an algorithm that is not at all or not very transparent (a numerical AI system) into a symbolic (and therefore more transparent) one, moreover already designed to show plastically the justifications and reasons for the inferences made.

Reasons can be of (at least) two types: explanatory and justification. From the logic of operation must therefore be distinguished the logic of explanation [*explanation*],[27] and the logic of justification [*justification*].[28]

than in terms of the existence of certain algorithms. The reason for this definition is theoretical: that is, it serves the purpose of obtaining a more general theory of *learnability*, since it is devoid of computational elements.

[27] For example, see, most recently, SCHNIEDER (2011).

[28] For example, see, recently, ARTEMOV and FITTING (2021) and FAROLDI 2019.

A case study.

I introduce my proposal of a justification logic for algorithm transparency with a case study on deep reinforcement learning.
While there are algorithms that by their nature are transparent, there are many that are opaque, in the most pregnant sense seen above.[29] As an illustrative case of an opaque algorithm, consider deep reinforcement learning, which has tremendous success and very little transparency.
In reinforcement learning (RL) in general, the aim is to associate situations with actions by maximizing a numerical reward function [*reward function*].
The agent tries and retries various actions, gradually refining its strategy depending on the reward (or lack thereof) it receives. In deep RL, a deep neural network is used, and the space of states is not explicitly provided.[30]
Juozapaitis et al (s.d.) propose decomposing the reward function so that the positive and negative factors of the decision are immediately visible.

The goal always remains maximizing the collective reward function, but now it is decomposed into some relevant types $c \in C$ with a vector function $\vec{R} : S \times A \rightarrow R^{|C|}$. Thus, it is possible to obtain the value of a stock with respect to only one

[29] In fact, there is no consensus on a general (or generic) method of explanation because of the great diversity of algorithms and methods in circulation. See HEUILLET et al (2020) in this regard.

[30] Precisely, a Markov decision process (MDP) is a tuple (S,A,T,R), where S and A are finite sets of states and actions, T (s, a, s′) is the probability of transitioning to state s′ after performing action a in s, and R(s,a) is the reward for performing a in s. A policy $\pi(s)$ returns an action to be taken in state s and the associated function Q, $Q\pi$ (s, a), provides the expected infinite-horizon γ-discounted cumulative reward of performing an action a in state s and then following π. By $Q*$ (s, a) we denote the function Q of the optimal policy $\pi*$, which satisfies the equation $\pi*(s) = \arg\max_{a \in A} Q*(s, a)$. Although standard, this formulation is taken from Juozapaitis et al (s.d.).

relevant factor: $Q\pi\ (s, a) = \sum_c Q\pi c\ (s, a)$, where the function Q returns the comulative reward.[31]

This decomposition is sufficient, according to the authors to explain why an action a1 is preferred to an action a2, in state s (i.e., $Q(s, a1) > Q(s, a2)$):

> we define the reward difference explanation (RDX) as the difference of the decomposed Q-vectors $\Delta(s, a1, a2) = \vec{Q}(s, a1) - \vec{Q}(s, a2)$. Each component $\Delta c(s, a1, a2)$ of the RDX is a positive or negative reason[s] for the preference depending on whether a1 has an advantage (disadvantage) over a2 with respect to reward type c (Juozapaitis et al (s.d.)).

In other words, the difference in reward for each relevant factor, which is a numerical assignment, is qualitatively regarded as a reason in favor of that action (relative to another) if the difference is positive, a reason against if the difference is negative.

Now we can use the logic of reasons (cf FAROLDI 2019) to reason with reasons. Each component $\Delta c(s, a1, a2)$ becomes a t-term that configures a reason for an utterance or action $\boldsymbol{\varphi}$ or $\sim\boldsymbol{\varphi}$ depending on whether it is a positive or negative reason.

So we have a language, totally readable and interpretable by a human being, that is able to aggregate reasons (t + s), combine them (t.s), and study their interaction with what they support to arrive at conclusions.

An example is as follows:

t : A; premise

c : (A -> A v B); classical logic

(c.t) : A v B principle of application, modus ponens

[31] This implies, of course, that the function is separable, which is not necessarily a given in normative contexts. See FAROLDI (2019) and FAROLDI (2021) for a discussion along these lines.

Each step or its conclusion (A, A -> A v B, A v B) has a reason beside it (*t*, *c*, *c.t*) that justifies and explains it, even in a combined way (as shown by the combined reason *c.t*).

Many of these logical systems are correct and complete, and some are even decidable.[32]

The details of this proposal are obviously for work of a more technical nature. If there were no insurmountable obstacles, this proposal would provide an automatic technique for transforming an algorithm that is not at all or not very transparent (a numerical AI system) into a symbolic (and therefore more transparent) one, moreover already designed to show plastically the justifications and reasons for the inferences made.

This would be sufficient to begin to meet the transparency and explainability requirements of the proposed EU regulation, especially as explicated by the criteria of the recent Supreme Court order under consideration.

One may ask at this point whether the proposed solution is not worse than the problem, replacing knowledge of computer science with knowledge of logic. It seems to me that it is not: to fully fulfill the purposes of ex ante disclosure (as specified, for example, by Articles 13 and 14 of the GDPR), it is in fact not necessary for the subject to have advanced knowledge of logic. In fact, all that is required is a simple exposition of the semantics of the logic obtained, associating each symbol with its informal meaning. This last step, if feasible, would make the system comprehensible in relevant aspects even to the ordinary citizen, in the spirit of the legal requirements that are increasingly becoming established. Moreover, the fundamental difference between a logical system such as the one proposed and a numerical one such as the one under consideration is highlighted by one of the senses of transparency we have highlighted: the latter would not be intelligible in an absolute sense even to its creator.

[32] The execution logic may also be incomplete or undecidable without being incomprehensible.

Discussion and further development

It remains to be determined whether the procedure proposed in this work is generalizable. As a conclusion I present three possible problems with this approach, the discussion of which seems very promising for further developing the proposal of this work, both technically and substantively.

The first problem is this: one can legitimately ask whether these explanations are true reasons: the fact that one action is preferred over another because more advantage is gained by performing it (albeit relative to relevant factors) is not in itself a justification for preference, but simply a breakdown of the generally consequentialist purpose of this method (maximizing the reward function).

The second problem is that the relevant factors (those in set C) are specified ex ante, and not learned by the agent.

The third problem is whether the reasons in question are justificatory (value-based) or explanatory (metaphysical-epistemic) reasons, which may not necessarily coincide, either conceptually or in their logical behavior.

General AI and Transparency: Two Critical Points of the EU AI Act.

Abstract. This chapter puts forward two broad criticisms of the proposed EU AI Act. First, the proposal does not address the future developments of general intelligent systems and it is ill-equipped to deal with more general existing systems like GPT-3. Second, one suggestion found in the literature to meet the first criticism, i.e. to increase transparency, fails, and not only because the proposal itself fails to specify the transparency standards required, but because it is problematic to spell transparency out in a concrete, usable way.

Keywords: *General AI; Transparency; EU AI Act*

"*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.*"
Alan Turing, 1951

Introduction

In recent years, with the extremely rapid progress in the field of artificial intelligence, the possibility of *singularity* has come to the attention of both specialists and the general public: machines that, having reached the stage of artificial general intelligence (AGI), or superintelligence, would give rise to an "explosive" moment that could radically and irreversibly change the planet and human civilization (cf. Kurzweil 2005, Bostrom 2014, Russell 2019).[33]

Already some time ago, Good introduced the ideas of superintelligence and singularity in the following way:

> *"Let an ultraintelligent 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 ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus, the first ultra-intelligent*

[33] In Faroldi 2020 and 2021a, I argued that such agents should be included in our normative practices, e.g. responsibility attribution.

> *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. … It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make."*

This raises many issues, but we focus, for the purposes of this paper, on the so-called *control problem*: how can we make sure that general intelligent, or so-called superintelligent, agents do not take control over us and the world?[34]

Section 2 argues that the proposed EU AI Act fails to address AI systems that do not have a concrete intended application. Therefore, (i) it fails to properly take into account the control problem due to potential general AI; (ii) even more modestly, fails to account for existing non-specific systems, like GPT-3.

Section 3 argues that one suggestion found in the literature to meet the first criticism, i.e., to increase transparency, fails, and not only because the proposal itself fails to specify the transparency standards required, but because it is problematic to spell transparency out in a concrete, usable way.

General AI and the control problem in the EU AI Act.

There are several solutions that have been proposed to deal with the control problem. The most popular is alignment: superintelligent AI must be programmed to align with human values.[35]

The proposed EU AI Act does little to nothing to address the control problem, as I will argue in a moment.[36]

This could be because the Commission does not want to overregulate a booming sector, and hinder future technological development.[37]

---

[34] Assuming we want that, of course. It is (at least theoretically) imaginable, in fact, that humanity could want to defer and give control to a superintelligence.

[35] Cf. Russell 2019, Christian 2020. As is easy to imagine, this proposed solution raises perhaps more problems than it promises to solve, from the problem of identifying and formulating human values, to the problem of instilling them in the AI systems in question. A second solution proposes to limit the capabilities of a AGI by isolating it from the outside world. A third solution proposes instead to increase the capabilities of humans in various ways to be on par with superintelligent systems. Cf. Ngo 2020.

[36] As a reminder, albeit a bit roughly: At the EU level, a regulation, once adopted is immediately valid, thereby obtaining uniformity in all EU countries; a directive, once adopted, is not immediately executive, as each country has to introduce it into its legislation in a potentially different way.

[37] "[the] proposal presents a balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining

A more balanced approach could go in many ways, but it has to start from recognizing the problem in the first place. AI risk researchers are mostly in agreement that alignment has to be built in before a singularity is reached,[38] i.e. before it is too late for humans to gain back control.

However, the Commission writes that

> *"[t]he proposal sets a robust and flexible legal framework. […] it is comprehensive and future-proof in its fundamental regulatory choices (AI Act, Explanatory Memorandum)."*

This thought is probably based on the guiding idea that

> *"legal intervention is tailored to those concrete situations where there is a justified cause for concern or where such concern can reasonably be anticipated in the near future (ib.)."*

Given the current technological development, general AI cannot be anticipated to happen in the near future, i.e. within the next 5-10 years. But come what may, it does not matter, it seems, because:

> *"At the same time, the legal framework includes flexible mechanisms that enable it to be dynamically adapted as the technology evolves and new concerning situations emerge (ib)."*

So the proposed regulation will run after and follow, rather than shape, technological development, which seems to be a grave mistake when it comes to general AI and the control problem. The shared idea is that, when the singularity is reached, it will be too late to do anything about control.[39]

An AI system is considered high-risk if it is intended to be used in a specific way,[40] either listed in Art 6 or in Annex III.[41] Notable examples of high-risk systems include:

- AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;

---

or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market (AI Act, Explanatory Memorandum)".

[38] Cf Christian 2020, Ngo 2020.

[39] As an example, cf. Ord (2020) and Ngo (2020).

[40] "The classification of an AI system as high-risk is based on the intended purpose of the AI system, in line with existing product safety legislation. Therefore, the classification as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used (EU AI Act)."

[41] "AI systems intended to be used as safety component of products that are subject to third party ex-ante conformity assessment; other stand-alone AI systems with mainly fundamental rights implications that are explicitly listed in Annex III [which] contains a limited number of AI systems whose risks have already materialised or are likely to materialise in the near future (EU AI Act)."

- AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;
- AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions;
- AI systems intended to be used for recruitment or selection of natural persons; AI intended to be used for making decisions on promotion and termination of work-related contractual relationships;
- AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services; AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score;
- AI systems intended to be used to dispatch emergency first response services;
- AI systems intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for crimes or the risk for potential victims of criminal offences;
- those with applications in migration, asylum and border control management;
- those with applications in the administration of justice and democratic processes.

We see that what matters in classifying an AI system as high-risk is its intended application in a field that is considered particularly relevant or worth extra care, and not at all the intrinsic characteristics of the AI system in question.

For instance, this proposed regulation could result in a low-level AI system used to categorize job application files in an alphabetical order as high-risk, and a superintelligent system with no concrete application or specific intended use, which might go on and influence millions of users in unforeseen, uncontrolled ways, as not high-risk.

This is problematic, because it seems to exclude currently existing systems that do not have a specific, concrete intended application, but are fairly general (without being instances of a general AI), such as GPT-3 and others.[42]

GPT-3 (short for Generative Pre-trained Transformer 3), created by Open AI, is a deep learning autoregressive language model that generates human-like text. Originally released in mid-2020, GPT-3 was trained on almost half a trillion tokens, is capable of coding in CSS, JSX, Python, and has been exclusively licensed by Microsoft.

[42] It is worth reminding here the definition of 'intended purpose' of Art. 3: "'intended purpose' means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation".

David Chalmers, a prominent philosopher, described it as "one of the most interesting and important AI systems ever produced",[43] but the MIT Technology Review says that its "comprehension of the world is often seriously off, which means you can never really trust what it says".[44] Other critics described it as unsafe, citing sexist, racist, and other biases. In one health-care simulation, GPT-3 advised a mental health patient to commit suicide.[45]

Regardless of this, GPT-3 is a system without an intended specific use (in the sense specified above): instead, it can be used in many, ways, unforeseen by its creators. Thus, it seems it cannot be classified as high-risk according to the EU proposal.

But suppose that it is used with a particular application in mind, which will not result in having it classified as a high-risk system.

This does not change the open nature of the system, which has now slipped regulatory measures.

That the Commission does not foresee AGI systems is also revealed by the provisions on human oversight, which require that a human:

> *"(d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;*
>
> *(e) be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure. (Art 14(4)).*"

These provisions do not take into account that an AGI could presumably (i) take control of its functioning by disabling and preventing any way to be switched off; (ii) make sure that human controllers are either fooled or convinced that everything is alright.

When it comes to (i), consider that even a non-malignant AGI, if built as many of the current AI systems, will have to maximize some reward function. Being switched off is definitely one way not to maximize the reward function, and therefore something to be avoided and prevented. Doing this with the full (supposed) power of an AGI is imaginably easy. When it comes to (ii), even without having reached a state of general intelligence, current algorithms are able to influence large numbers of people (e.g. through making fake news viral).

To these criticisms of the EU proposed regulations (and in particular of Art 14 in the present context), one can object that they are based on mere projections, or extrapolations, based on the capabilities of current AI systems, with no guarantee of what might or might

[43] https://dailynous.com/2020/07/30/philosophers-gpt-3/#chalmers

[44] https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion

[45] https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/

not happen in the future. While this is true, one should also consider that the moment an AGI is created and is allowed to take control, there is a very high chance that it will not be possible to go back: once a singularity is reached, there won't be the possibility to switch off the systems in question. While there are many disanalogies, it is plausible to liken the singularity with nuclear annihilation: a point of no return. Seen in this light, it makes sense to legislate already now to safeguard and prevent this possibility, even if remote, rather than complain later, when too late.

But how to do that?

One proposal to address a similar concern suggests to widen Title IV, on transparency obligations, to apply across all AI applications regardless of specific purpose.[46] One way this could be achieved, according to the authors, is by requiring a "complete risk assessment of all an AI system's intended uses (and foreseeable misuses (ib.))".

However, this is also problematic: transparency is a notoriously complicated requirement, and the EU AI act proposal does not do anything to improve the situation. Let's see this aspect in turn.

Transparency in the EU AI Act.

The EU proposed AI Act imposes some transparency requirements to certain systems:

> *"Transparency obligations will apply for systems that (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content ('deep fakes')."*

While transparency is not at all a clear concept in the literature, it has to do with knowing how a system works, at least at a high level, to know how decisions are taken, for instance.[47] This seems also to be taken up in some existing legal sources.[48]

As a first point to notice, transparency obligations are not imposed on the same systems that are classified as high risk (see above).[49] Therefore, we can already exclude that more transparency is sufficient, *per se*, to solve the problem put forward in the previous section.

[46] FLI position paper on the EU AI act, 2021 https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665546_en

[47] This characterization is obviously very simple-minded.

[48] In Faroldi 2021b, I put forward critical remarks when it comes to transparency requirements, also with regard to the GDPR and a recent order of the Italian Corte di Cassazione. The GDPR prohibits automated decisions that have consequences for individuals, and establishes a right to meaningful information about the logic employed in these procedures. A recent Corte di Cassazione order (n. 14381/21, May 25, 2021) places emphasis on access to the elements an algorithm uses in a decision and the executive scheme in which the algorithm in question expresses itself.

[49] At least prima facie. It is possible that these two different fields of application end up covering exactly the same systems, although this is extremely unlikely.

If it were, in fact, it would be enough to require transparency to mitigate the potential problems of high-risk systems, without the need for further measures.

However, there's a further difficulty. What the EU Commission means with 'transparency' does not really seem aligned with the rest of the literature. The EU Commission seems to have in mind that transparency consists in the right to be told that one is interacting with an AI system:

> *"When persons interact with an AI system or their emotions or characteristics are recognised through automated means, people must be informed of that circumstance. If an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, there should be an obligation to disclose that the content is generated through automated means, subject to exceptions for legitimate purposes (law enforcement, freedom of expression). This allows persons to make informed choices or step back from a given situation."*

Fortunately, the EU AI Act also requires some form of control and "transparency", again for high-risk systems, in Art. 13, which is glossed in the following way:

> *"To address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate (EU AI Act, Preamble, 47)"*

The fact that Art. 14 requires human oversight, in such a way that the human in question:

> *"fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible (Art 14 (4)(a))"*

tells us that the certain degree of transparency (in the non-opacity sense) required for users is lower than that required for the controllers.

But is the transparency standard required for controllers even possible? Before answering this important question, it is necessary to check what it is meant with 'transparency' and its antonym, 'opacity', in the literature.

One can distinguish at least three senses of opacity, i.e. not transparency.

ALREADY DONE

There is *intentional* opacity when the algorithm is not voluntarily made public in order to maintain a competitive or economic advantage. This type of opacity could be "solved" by simply publishing the algorithms in question - except for subsequent problems with competitiveness, etc.

There is *cognitive* opacity when the algorithm, while transparent, simply cannot be interpreted by most users due to their ignorance. In this case, moving to an open system, the opacity is not resolved ipso facto: an intermediate category of interpreters, explanators, or massive investments on the education of the population are needed.

There is *intrinsic or essential* opacity when it is the algorithms themselves, regardless of programmers and publication, that make decisions in a way that cannot be explained or interpreted by a human being, also regardless of cognitive capacity. This is the case with some machine learning techniques, and deep neural networks.[50]

It is of course this latter sense that is relevant, and it is often investigated under the label of 'interpretability'.

Lipton (2016) gives a systematic account of the relevant literature, and argues that the interpretability of models falls into two broad categories: that of transparency *stricto sensu* and that of *post hoc* explanations. In the first, what matters is understanding the mechanism by which the model works. In the second, what matters is extracting information from the models to clarify what exactly they learned.

Krishnan (2020) however, argues more radically that the concept of 'interpretability', is vague: thus, it is very hard to know whether a given technical solution is acceptable or not; and in any case often that of 'interpretability' is a means to achieve other ends (such as non-discrimination or justification). We should therefore not require algorithms to be interpretable, but focus the discussion on the true ends we want to achieve.

At a very high level, transparency should entail that we are in a position to know why a certain algorithm produced a certain output. The first 'why' we can distinguish is that of an explanation, which is of course a whole problem in itself. At a minimum, it should be informative about the causes and the processes that lead to that output, given that input, at an appropriate level of granularity.

The second 'why' we can distinguish is that of a justification, e.g. about the reasons in favor (or against) that particular output or set-up.

There is a definite sense in which this second 'why', that of justification, is more normative than the first.

[50] For a similar tripartition, see Burrell (2016). There are also algorithms that are totally transparent because of their architecture, which makes them self-explanatory, such as linear regression, decision trees or rule-based systems.

The confusion between these two "why's" is already apparent in some technical proposals of explainability in machine learning.[51]

Given the multiple realizability and vagueness, it is unclear how these notions of transparency found in the technical literature can harmonize with what is or will be required in artificial intelligence legislation.[52]

Conclusion

In this paper I argued that there are (at least) two problematic points of the proposed EU AI Act: first, it is ill-equipped to deal with more general, already existing systems like GPT-3 and it is even in worse position to deal with possible strong or general artificial intelligent systems; second, I argued that one suggestion found in the literature, i.e. to increase transparency, fails, and not only because the proposal itself fails to specify the transparency standards required, but because it is problematic to lay them down in a concrete, usable way.

One recommendation that emerges from this short paper adheres to the principle that, when it comes to strong AI, the EU AI Act should lead and shape technological development, rather than just follow it, because once (if) a singularity is reached, it will be too late. The recommendation is to take into account already *now* the possibility of strong AI in the *future*. Much more work is needed in setting out what is to be done concretely, which is an active field of research in AI safety, but awareness and risk identification is a first step that can and should be already embedded in the EU AI Act.

[51] Cf. e.g. Juozapaitis et al (s.d.)., which I discuss at length in Faroldi 2021b.

[52] A couple of recent high-level approaches suggest to give more prominence to symbolic techniques in AI, especially when it concerns certain values (like transparency) in a legal context. Cf Billi (2021) and Faroldi (2021b).

Philosophical considerations on the status of superintelligent artificial agents

This essay focuses on the issue of the treatment of superintelligent artificial agents, proposing and investigating the hypothesis that superintelligent agents can enter fully into the attribution of responsibility (moral, criminal). The argument is based on a proposal to reevaluate the subjective elements with regard to the attribution of criminal responsibility.

Introduction

Artificial intelligence (AI) consists of systems (software or hardware) that function by developing or possessing characteristics that, in humans, correspond to higher cognitive functions, such as thinking, problem-solving, and learning.
It is well known that many of today's AI systems can develop problems that in some cases have regulatory, ethical, or legal implications, due in large part to contingent limitations: think of flawed algorithms that introduce racial bias, lack of transparency in decisions, and so on.
But there are also problems with ethical and legal implications due to the fact that AI systems work too well: think of near-instantaneous facial recognition that can lead to Orwellian surveillance systems, the accuracy of *deep fakes* videos that can create social unrest, the likely replacement of most low-skilled jobs in the coming decades with a potential massive increase in the unemployment rate, the status of robots being used as emotional or sexual companions, or as *caregivers*.[53]

Without indulging in millenarian considerations, intelligent agents (such as *self-driving cars*) now interface with different parts of society on a daily basis, exhibiting an increasing degree of 'autonomy', and less and less connection to the original creators. But what is meant by 'intelligent'? What is the relationship between intelligent non-human agents and responsibility?

[53] For such a list, see LIAO 2020.

In recent years, with the extremely rapid advances in the field of artificial intelligence, the possibility of the singularity has come to the attention of both specialists and the general public: machines that, having reached the stage of general artificial intelligence (or superintelligence), would give rise to an "explosive" moment that could radically and irreversibly change the planet and human civilization (cf. Kurzwell, 2005, Bostrom 2014, Russell 2019).

I distinguish two questions at this point: What is (what would be, if they existed) the status (legal, ethical, or more generally normative) of these super-intelligent agents? What should (should) be their status?
In the rest of this essay I will focus on the *second* question, which can be declined into two basic questions: how should we treat intelligent or superintelligent artificial agents? How can we ensure that they treat us as we want to be treated?

The second issue, also referred to as the *control problem*, has several opposing solution proposals. The most popular is alignment [*alignment*]: superintelligent AI must be programmed to align with human values.[54] A second solution proposes to limit the capabilities of a super AI by isolating it from the outside world. A third solution, on the other hand, proposes increasing the capabilities of humans in various ways to be on par with superintelligent systems.

Instead, this essay focuses on the first question, proposing and investigating the hypothesis that superintelligent agents can enter fully into the attribution of (moral, criminal) responsibility. The argument is based on a proposal to reevaluate the subjective elements with regard to the attribution of criminal responsibility, a proposal elaborated in Faroldi 2020 that makes essential use of the concept of reason [*reason*].[55]

[54] As is easy to imagine, this proposed solution raises perhaps more problems than it promises to solve, from the problem of identifying and formulating human values to the problem of instilling them in the AI systems in question. As one reviewer points out, programming (design) is one thing, the use of this agent is another, and autonomous (or conscious) decision-making is another, among the various distinctions that can be advanced in this contensto.

[55] This essay in fact picks up and develops some of the themes addressed in Chapter 7 of FAROLDI 2020, on which it is partly based.

Responsibility , Intelligence, Superintelligence

The definitions of accountability, intelligence and artificial intelligence present obvious difficulties, which are, however, due to different reasons and can be at least partially overcome by adopting diversified strategies.

Accountability inquiry can take at least *two* forms: prescriptive and descriptive (cf. Faroldi 2014).
With regard to both intelligence and artificial intelligence, the second path is precluded; we will therefore have to rely on a (moderately) *prescriptive* perspective, albeit an argued one that takes as a critical starting point the consensus of scholars in the respective disciplines of psychology, psychometrics, and the artificial intelligence and machine ethics community.

An exact definition of intelligence is problematic. First, one must distinguish intelligence from all kinds of consciousness and awareness: not necessarily an intelligent being is also endowed with those mechanisms that in humans identify with higher functions such as consciousness or awareness. This consideration seems extremely relevant when we go to consider the ramifications of intelligent nonhuman agents particularly in the areas of morality and law.
Second, we might be tempted to follow a strategy proposed by the English mathematician and logician Alan Turing, who in (Turing 1950) proposed replacing a direct test of intelligence with an indirect test based on the interpretation of written responses (called the imitation game), with the aim of ascertaining whether a machine is capable of thinking. Given the difficulty of defining what thinking is, Turing's strategy falls into the rut of those that substitute for the assessment of intelligence a trait that (according to Turing) would be inseparable from intelligence itself (though not coinciding with it): imitation and interpretation of language.
In short, one person interprets written answers that are submitted to him by two other people in relation to questions that he imposes. Beyond the details, what matters is that at a certain point one of the two people providing answers is replaced by the machine. If the first person still fails to notice anything, the machine has passed the test and can thus be considered intelligent. The so-called Turing test is not all that innovative: as Turing himself acknowledges, it is based on an insight of Descartes in the *Discours de la méthode* and has been taken up several times throughout the history of philosophy by, for example, Diderot

and Ayer, though in less developed and direct forms. Equally well known perhaps is the mental counter-experiment proposed by the American philosopher John Searle, known by the name of the Chinese room argument. However, it is well known that the Turing test has not been passed by any machine, at least in most versions of the test known to me.

Third, it should be considered that most psychologists now agree that in addition to specific cognitive abilities there is a general factor of intelligence, called *g*, which is related to most tests for measuring intelligence, either directly or indirectly. [56]

The idea of the existence of a general intelligence factor is perhaps the key to the issue. Indeed, there is no doubt that for well-defined, plausibly computable tasks, so-called machines outperform humans not only on average, but also among the samples the experts in those well-defined tasks. Examples can be counted numerous even limiting oneself to the 10s in the current century: one for all is worth the crushing victory of Google's artificial intelligence AlphaGo over the multiple-time world champion of go, an extremely complex game, Lee Sidol.

We would not call AlphaGo intelligent, however, given its general incompetence to tackle any kind of task for which it has not been preprogrammed (cf. Russell 2019, Ch. 3). For the ability to tackle novel tasks that are not strictly preprogrammed, we speak of *general artificial intelligence*. In the course of this chapter we will take the following consideration as the definition of work as a mere starting point of super intelligence:

> Let an ultraintelligent 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 ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent 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 (Good, 1965).

[56] Cf. STERNBERG & KAUFMAN (2011).

Analysis

Assume then, that there is a non-human agent endowed with intelligence similar to or superior to human intelligence in relevant respects.
These agents will be autonomous in the sense that they will not necessarily adopt a narrow set of behaviors pre-programmed by the creator; rather, they will manifest characteristics of autonomy of both means (to achieve a plausibly pre-programmed outcome) and ends.[57]

Preliminary issues

There are several perspectives open at this point.

A preliminary issue is the cognitive architecture of these agents in relation to regulatory orders. This issue begins to be addressed, albeit only partially and no doubt without autonomy and "superintelligence," in the 1910s in relation to so-called *self-driving cars*.[58]
Reasonably, normative systems are codified as a set of closed utterances under inference rules, so that by taking in information from outside, these systems can categorize and act with the normative systems of reference in mind.[59]
First, a serious problem is that normative systems are written in natural language, and therefore a translation is needed. Beyond technical issues, there are at least two well-known issues that both mathematical logic and legal and deontic logic highlight that go beyond the mere difficulty of translation into formal language. First, very reasonable to think that, to adequately translate normative systems into formal language, one needs at least all the expressive capacity of second-order logic to quantify on properties and relations. This implies that using the most natural semantics for second-order logic is necessarily

[57] In the machine ethics community there is, of course, an extensive literature on the subject. For an influential starting point, see: MOOR 2006.

[58] The Report "Liability for Artificial Intelligence (and other emerging technologies)," prepared by the New Technologies formation group of experts established at the European Commission, then adopted in February 2020, for example, does not deviate much from conservative positions in terms of legal personality and damages.

[59] 'Taking into account,' in this context, is used to refer to a broad interaction with regulatory systems that goes beyond classical interactions (compliance, default, etc.). As one reviewer points out, in this regard it is interesting to ask whether systems that already use pre-specified rules can become super-intelligent. See WINFIELD et al, 2014.

incomplete.[60] In colloquial terms, correct logic is incomplete if the computational system never derives falsehoods, but cannot derive all truths: this means that there are truths inaccessible through the computational system.

Second, it is well known that standard deontic logic is (i) inadequate to render all the complexities of normative language, given that the analysis is still in progress and a consensus has not been reached; (ii) full of paradoxes and puzzles with counterintuitive consequences, which an "illogical" human agent would never commit.[61]

In general, it is reasonable to assume that by the time there is a superhuman artificial intelligence (if there is one) a way will also have been found to resolve these uncertainties. It should be noted, however, that it is unclear to what extent an intelligent agent can "reprogram" itself, and thus change the type of reasoning it uses, the rules of inference, and the way it learns and understands.

Substantive issues

Let us now turn to less technical and more substantive issues related to the relationship between superintelligent autonomous agents and criminal law.[62]

First, an ontological question: in what category (legal, for the purposes of this essay) should these autonomous agents be situated? Plausibly they have traits that draw them relevantly close to human beings, but they lack other (more biological and emotional) traits that human and non-human animals have in common, and which are often used as justifications for the attribution of sanctions.[63]

[60] There are a number of possible answers to this problem, but they always fall back on using first-order logic, either directly or by simulating higher-order entities through *multi-sortal* first-order logic. First-order logic is known to be complete and semidecidable.

[61] Recent attempts to develop deontic logic that is suitable for agents with limited abilities and solves some of the paradoxes and puzzles can be found in Anglberger et al. 2016, Faroldi 2019.

[62] On the question of intelligent (not superintelligent) agents in law (and in morality) there is beginning to be a reference literature: see: Floridi and Sanders (2004), Chopra and White (2011) and Asaro (2015), among others. For issues both general and in Italian law, cf. Alpa 2020 and Dorigo 2020. Much of the legal literature, however, does not consider the discontinuity represented by a *general* artificial intelligence, or superintelligence.

[63] But note that there is literature on the subject of testing the existence of "consciousness" in artificial intelligent agents. For a review, see Schneider (2020).

Second, one may ask what protections should be extended to these agents. If it is true that they have certain characteristics of autonomy, initiative, and autonomous ends, in part comparable to human beings, it is reasonable to assume that certain protections should be afforded, radically differentiating non-autonomous agents (for whom a *kill-switch* could without qualms be admitted) and autonomous agents, for whom it seems unreasonable not to admit a judicial proceeding (if not for the protections of the agents themselves, for the interests of human beings who might in some way depend on them, materially or emotionally).

Third, we may ask how and whether we assign responsibility for the "actions" of intelligent autonomous agents.[64]

It is well known that the Italian republican charter prescribes that criminal liability is personal (art 27 co. 1). For a long time, and based on a literal reading of the constitutional dictate, the only exclusion was that of liability for the acts of others, and indeed not that of cases of strict liability for one's own act, the mere "material causal link" being sufficient in this case.

That was until the fundamental (and highly debatable) Constitutional Court ruling No. 364/1988 declaring unconstitutional Article 5 cp in the part that does not provide for inexcusability of unavoidable ignorance of the criminal law. Beyond the specific subject of the ruling, the judgment is extremely relevant because it recognizes a constitutional guarantee to the so-called principle of guilt:

> one is criminally liable only for one's own fact, provided that it is made clear that by "one's own fact" is meant not the fact linked to the subject, to the author's action, by the mere link of material causality [...] but also, and above all, by the subjective moment, constituted, in the presence of the foreseeability and avoidability of the prohibited result, at least by "guilt" in the strict sense.

[64] And what sanctions can be expected. See on this topic HALLEVY (2015).

Thus the so-called principle of subjectivity is recognized, which requires that there be a "psychic" connection between the perpetrator and the conduct itself. Even sharper is the Court's position in Judgment 1085/1988 against strict liability:[65]

> In order for Article 27, first paragraph, Const, to be fully respected and criminal liability to be authentically personal, it is essential that each and every one of the elements that combine to mark the disvalue of the case must be subjectively linked to the agent (i.e., be invested with malice or guilt).

Of note, finally, is the important (and debated) unified sec. sentence Cass. Jan. 22, 2009, no. 22676, so-called Ronci sentence, in which the supreme court, in relation to art. 586 of the Criminal Code (Death or injury as a consequence of another crime) definitively sides against the notion of strict liability and in favor of fault in unlawful activity, fully completing the path that has seen the affirmation of the constitutional rank of the principle of culpability.

---

[65] Compared to other legal systems, which are more generous in providing for various cases of strict liability, the Italian legal system has very few and residual ones. One of the best known, albeit for the actions of others, was certainly the original formulation of Article 57 of the Italian Criminal Code ("in the case of a periodical press, whoever holds the position of editor or responsible editor is liable for the crime committed, without prejudice to the liability of the author of the publication"), which was updated in a culpable omission sense by Law 127/1958. We then note Article 609 sexies, which instituted a case of strict liability in the case of certain crimes against minors until 2012, when it was tempered in the sense of the exclusion of unavoidable ignorance by Article 4 of l. Oct. 1, 2012, no. 172 also following the *judgment C. cost. 24.07.2007, No. 322,* which affirmed the need for an interpretation in accordance with the principle of culpability.
Contrary, on the other hand, and in favor of strict liability in relation to the age of the offended person in sexual assault offenses (provided for in Section V of the *Sexual Offences Act* of 2003), is the judgment C. eur. dir. uomo, sec. IV, dec. Aug. 30, 2011, app. no. 37334/08, G. v. United Kingdom. I do not know of any discussions in the literature, but it seems to me that some of the offenses under Article 1 of l. 205/1993 (the so-called Mancino law) count in this sense. Specifically, "Any organization, association, movement or group whose purposes include incitement to discrimination or violence on racial, ethnic, national or religious grounds is prohibited. Whoever participates in such organizations, associations, movements or groups, or assists in their activities, shall be punished, for the mere fact of participation or assistance, by imprisonment of six months to four years. Those who promote or direct such organizations, associations, movements or groups shall be punished, for that alone, by imprisonment of one to six years."

Some authors hope for the definitive extension even in the remaining cases that can be constructed as remnants of strict liability.[66]

Excluding therefore the possibility of strict liability in the current legal system, we can move on to consider the liability of entities for administrative offenses arising from crimes (so-called criminal liability of entities), introduced in the last two decades in most major legal systems, and notably in Italy by Legislative Decree 231/2001.

The comparison with the criminal liability profiles of entities is relevant in some respects, irrelevant in others.

In particular, it is relevant because profiles of criminal liability are attributed to individuals who are not inherently endowed with those subjective elements usually considered necessary for criminal liability:[67] this argues in favor of the thesis that criminal liability is attributable even without reference to a particular psychological state (because it is contingently or necessarily absent).

It comparison, however, is irrelevant in another respect: someone who helped commit the acts in question is a natural person who possesses those psychological requirements ordinarily required: a fundamental difference from the case of liability of (super-)intelligent agents. Mitigating this consideration of irrelevance cannot be the objection that, under Article 8(1)(a), the liability of the entity remains even when the perpetrator of the crime has not been identified or is not imputable: in fact, the perpetrator (or perpetrators) who is a natural person of the crime has been there anyway, regardless of the procedural condition.[68,69]

[66] On this topic I point out, for all, BASILE (2011).

[67] While recording, in the literature, the thesis that collective entities have a collective consciousness.

[68] SOLAIMAN (2017) is perhaps the most up-to-date and comprehensive treatment of the issue of the legal personality of robots, beyond their intelligence or super-intelligence. The author takes a stand against the argument that they should be accorded legal personality by analogy to businesses or entities.

[69] Also interesting but irrelevant for the purposes of the issue addressed in this paper is the resolution passed by the European Parliament on February 16, 2017 on "Civil Rules on Robotics." Interesting because it sharply acknowledges the problems emerging from the new category of autonomous and intelligent non-human agents, irrelevant because it is limited to considerations of mere civil liability.

Even the case of the autonomous agent who is used as a mere means of executing the criminal will of the creator or principal poses particular problems for the existing normative set-up, not with regard to liability, in the hands certainly of the principal or creator, autonomy of the non-human agent aside, but for the diapraxic act of exploiting an autonomous agent for criminal purposes, thus adding another case in point (cf. also art. 111 cp, albeit with the limitation of the principle of legality and the prohibition of analogy).[70]

The logical space of possibilities is saturated by the consideration of the spectrum of cases from (a) the use of the autonomous agent as a mere perpetrator, who nevertheless retains some degree of independence, to (b) pretermitted conduct, to (c) criminal but completely autonomous conduct by the non-human agent.[71]

With regard to case (c), it is clear that, according to Article 85 of the Criminal Code, a person who lacks the capacity to be of sound mind is not chargeable.
However, having stipulated a certain definition of a superintelligent agent above, the capacities of mind and will are, albeit in a form not identical to that of human agents, reasonably and substantially possessed by the non-human agents in question.[72]

Perhaps the most interesting comparison is with *group agency*, that is, the phenomenon whereby organized sets of individuals can constitute an agent endowed with purposes *over and above* the individual members of the group. These collective agents are ascribed desires, beliefs, etc., and the ability to interact with the environment, even to the point of ascribing to them limited versions of legal personality distinct from that of their constituent members. Certain notions of collective responsibility apply to these groups. Christian List (2019) argues that these characteristics (beliefs, desires, interaction with the environment circumstances) are also shared with intelligent (non-human) agents: indeed, that *group agents* are neither more nor less than a specific case of an artificial intelligence system. Moreover, he argues that the responsibility of a *group agent* cannot be reduced to the responsibility of the individuals in it. Likewise, the responsibility of an artificial intelligent

[70] On the notion of a diapraxic act, see CONTE., 2012.

[71] More accurately: conduct that would be criminal in the current system if enacted by a human agent.

[72] At this point, as an anonymous reviewer suggests, we might ask why use, adapt, extend, or proceed by analogy with respect to a penal category such as imputability, if the capacity to understand and will not be identical to human capacity? Why not introduce a specific one? An acceptable answer would take us very far, and would necessarily have to address the purpose and nature of normative systems such as law. In this more limited context, I point to needs for system coherence and adherence toward pre- and extra-legal practices as reasonable motivations for such an approach.

agent cannot be reduced to the responsibility of an individual (human). Along with the requirement of knowledge and control, List believes that to be *fit to be held* responsible [*fit to be held responsible*] also requires moral agency [*moral agency*]. In the animal kingdom, for example, humans are the only ones who have all these requirements, and therefore the only ones who can be held responsible, List argues. What is the difference? Drawing on an argument by Philip Pettit (2001), we can see that while with nonhuman animals we can interact causally but not normatively, with *group agents* we can. They have what Pettit calls *conversability*, that is, they are able to appreciate and respond to reasons [*reason*]. This would be, according to Pettit and List, the figure of *moral agency*.

It is possible for intelligent but non-human agents to be moral agents in this rather limited sense of responding to reasons. At least two possibilities for having "moral machines" are recognized in the literature: first, morality rules and decision-making processes are encoded (*recus*: pre-programmed) *ab initio*; second, autonomous agents are able to reason morally by self-instruction, e.g., by extracting moral *patterns* from large data sets with (*supervised*) *machine learning* techniques.[73,74]

We should not forget, however, that we can be content to ascribe criminal, or more generally legal, responsibility, and that this does not depend in an essential way on moral notions, as argued in Faroldi 2014, 2020. In conclusion, many of the thorniest questions, e.g., about the moral agency of non-human entities, can be seen as orthogonal to the ascription of responsibility for legal purposes. In Faroldi 2020 a theory of (not necessarily moral) responsibility based on reasons is argued and elaborated, which can be extended to non-human (super)intelligent entities. It is through this passage (which we cannot go into here) that a notion of even legal responsibility can (and perhaps should) be extended to super-intelligent artificial agents.

[73] It is, of course, questionable in the latter case whether there is room in this scenario for a conception of morality that is vaguely Kantian, and to what extent Hume's law can be respected.

[74] There is a *third* proposal in the literature, explored by CONTISSA et al (2017) regarding autonomous vehicles that could be equipped with an "ethical knob," a mechanism that would allow passengers to customize the moral strategy of the vehicle in question. The vehicle would then have to implement the chosen moral theory, while the manufacturer would be responsible for making the passengers' choice possible and executable by the vehicle.

This at least partially answers the second question posed in the introduction (What should (should) be the status of super-intelligent artificial agents?) by showing that it is possible to consider them responsible members of our normative communities.

Objections, alternative proposals and open questions

The most immediate objection is, of course, that any sanction attributable to non-human autonomous agents is not (and cannot be) so much a penalty as a security measure. The reasoning is as follows: having ascertained (to some degree) the social dangerousness of the autonomous non-human agent in question, it would, in the absence of its own subjective elements, not be imputable, and therefore not punishable. However, analogous to the case of the human agent who is not punishable because he is not imputable, it could instead be sanctioned with a "personal" security measure, and exclusively *pro futuro*.[75]

The objection is difficult to fully evaluate, since-as is well known-the distinction between punishment and security measure is increasingly blurred, at least in the Italian legal system. In general, however, one can respond by noting that a hypothetical autonomous non-human agent, who is at least, if not more, intelligent than the average human, is by definition a *rational* agent, although perhaps not perfectly rational due to limited (computational) resources. Since he is a perfectly rational agent, and since in the Italian legal system punishment also has a special-preventive function, for eminently consequentialist reasons the punishment, rather than the security measure, seems appropriate to influence his future conduct by reason of his past conduct.[76]

An alternative solution to the one proposed here, and in some ways a very interesting one, is that of the principle of personality of law, insofar as it can be conceivable in terms transcending mere private law.

[75] A prerequisite of the discussion is, of course, that the responsibility to any human operators, creators, owners and regulators is not sufficient to account for the phenomena we are concerned with in this context, or has been exhausted or unaccountable (the so-called *responsibility gap*).

[76] To the extent that fragile or injured subjects are considered, the special-preventive function is declined almost in terms of the ideology of treatment, toward the health care intake of the subject in question. However, the primary interest here is not toward these subjects.

The basic idea is as follows: since these (super)intelligent autonomous agents nevertheless have such characteristics, both of patience of agency, as to see them guaranteed at least a part of subjectivity, but on the other hand do not perhaps fully enjoy those characteristics necessary for imputability, whether for retributivist reasons, consequentialist reasons, or preventive reasons, it is appropriate to introduce a specific (criminal) law specific to them, to which only they are subject. This specific law would have the advantage of not undermining the guarantees and institutions of "human" criminal law while recognizing, for better or worse, that is, in terms of rights and duties, a specific personality to superintelligent autonomous agents.[77]

There are also well-known problems with this approach. First, there are the typical problems of the principle of personality of law, having to do with interactions between different *nationes*. Second, still classical but unexpectedly, hybrids and chimeras will have to be carefully considered, which will not simply be the product of "natural" crosses between a Lombard and a Roman, but "radical" modifications and improvements of the human being as we know it. What will it prevent in this case? Would it be fair to call a hybrid back to the higher standards of a superintelligent autonomous agent?

One final problem remains to be addressed, perhaps the most intangible and speculative. The biblical idea of the benevolent lawgiver empowering the inferior and the contractualist idea of the social compact among equals seem to come into crisis here. Indeed, by granting some autonomous non-human agents some kind of superintelligence, it becomes questionable to justify why it should be the non-superintelligent human agents who should impose a certain right on them, and not vice versa.[78]

[77] The 2017 "Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))" already recommends, with regard to civil law, "creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause."

[78] RUSSELL (2019, p. 127) raises two objections to this: the first regarding the diminution of human dignity, and the second regarding the possible creation of two classes of humans: an elite who create these intelligent machines, and everyone else who is controlled by them.

References

ALPA, G. (2020) a cura di, *Diritto e intelligenza artificiale*, Pacini, Pisa.

ANGLBERGER, A., FAROLDI, F. L. G., e KORBMACHER, J. *An Exact Truthmaker Semantics for Obligation and Permission*, in *Deontic Logic and Normative Systems*, a cura di O. Roy, A. Tamminga, e M. Willer, DEON16, College Publications, 2016, 16–31.

ASARO P. M., *The Liability Problem for Autonomous Artificial Agents* in AAAI Symposium on Ethical and Moral Considerations in Non-Human Agents, Stanford University, Stanford, CA, 2016, 191.

BASILE, F., *L'alternativa tra responsabilità oggettiva e colpa in attività illecita per l'imputazione della conseguenza ulteriore non voluta, alla luce della sentenza Ronci delle Sezioni Unite sull'art. 586 c.p.*, in *Rivista Italiana di Diritto e Procedura Penale*, 2011, vol. 54, n. 4, 911–968.

BOSTROM, N., *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

CHOPRA S. e WHITE L. F., *A Legal Theory for Autonomous Artificial Agents*, Chicago University Press, 2011.

CONTE, A. G., *Diapraxía*, in *Ontologia del normativo. Studi per Gaetano Carcaterra*, a cura di D. M. Cananzi e R. Righi, Giuffrè, 2012, 419-424.

CONTISSA, G., LAGIOIA, F. e SARTOR, G., *The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law*, in *Artificial Intelligence and Law,* 2017, 25, 365–378.

DORIGO, S. a cura di, *Il ragionamento giuridico nell'era dell'intelligenza artificiale*, Pacini, 2017.

FAROLDI, F.L.G., *The Normative Structure of Responsibility. Law, Language, Ethics*, College Publications, 2014.

FAROLDI, F.L.G., *Hyperintensionality and Normativity,* Springer, 2019.

FAROLDI, F.L.G. *Responsabilità e ragione.* Satura, 2020.

FLORIDI, L. e SANDERS, J. W., *On the Morality of Artificial Agents*, in *Minds and Machines*, 2004, vol. 14, 349–379.

GOOD, I. J., *Speculations Concerning the First Ultraintelligent Machine*, in *Advances in Computers*, 1965, vol. 6, n. 1.

HALLEVY, G., *Liability for Crimes Involving Artificial Intelligence Systems*, Springer, 2015.

KURZWEIL, K., *The Singularity is Near*, Viking Press, 2005.

LIAO, S. M., *A Short Introduction to the Ethics of Artificial Intelligence*, in LIAO, S. M. (ed.), *Ethics of Artificial Intelligence*, Oxford University Press, 2020, 1–44.

LIST, Ch., *Group Agency and Artificial Intelligence*, London School of Economics, working paper, 2019.

MOOR, J.H., The nature, importance, and difficulty of machine ethics, *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 18–21, 2006.

PETTIT, P., *A Theory of Freedom: From the Psychology to the Politics of Agency*. Polity and Oxford University Press, 2001.

RUSSELL, S., *Human Compatible*. Viking, 2019.

SCHNEIDER, S., *How to Catch an AI Zombie: Testing for Consciousness in Machines*, in LIAO, S. M. (ed.), *Ethics of Artificial Intelligence*, Oxford University Press, 2020, 439–458.

SOLAIMAN, S.M., *Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy*, *Artificial Intelligence and Law*, 2017, vol. 25, n. 2, 155-179.

STERNBERG, R. J., e KAUFMAN, S. B., a cura di, *The Cambridge Handbook of Intelligence*, 2011, Cambridge University Press.

TURING, A., *Computing Machinery and Intelligence*, *Mind*, 1950, vol. LIX, n. 236, 433-460.

A. F. T. WINFIELD, C. BLUM, e W. LIU, Towards an ethical robot: Internal models, consequences and ethical action selection, in: *Lecture Notes in Computer Science*, vol. 8717, Berlin, Germany:Springer, 2014, 85–96.

Delphi: Towards an Ethical AI? No.

FOR AN AI TO BEHAVE ETHICALLY, TRAINING DATA SHOULD NOT BE BASED ON WIDESPREAD PRACTICES

A deep-learning prototype of language-based common-sense moral reasoning called delphi (https://delphi.allenai.org/) shows much greater accuracy (as vetted by humans) than the zero-shot performance of GPT-3.[1] This could be a stepping stone in the important but extremely difficult research program of having machines learn human values.
In Delphi, one can type questions and get answers like "It's okay" or "You shouldn't do that". There are obvious minor problems, such as being subject to framing effects, but there are two major design problems that could lead the AI researcher community astray.

First, the use of descriptive ethics. Descriptive ethics is about the ethical beliefs and practices of a given community at a given time. It is an empirical matter. Normative or prescriptive ethics, instead, is about what is right to do, or what should be done. Humans are widely considered to be morally imperfect. Basing moral answers on what people do (and say they do) is a recipe for disaster.

Second, the mistake between norms and normal. There are practices that are normal, in a statistical sense. Think about slavery in Ancient Greece, or racism very recently, perhaps eating animals today. That they are normal may also result in people approving of them. However, this is not enough for them to be norms. Sticking to what is normal today will impede moral progress, be it of humans or artificial agents, in the future.

[1] Liwei Jiang et al., ***Delphi: Towards Machine Ethics and Norms,*** ArXiv, abs/2110.07574 (2021).

Notes on Russell's Human Compatible Approach

This chapter aims to contribute to the issue of embedding human values in artificial general intelligence (AGI), by suggesting the development a new theoretical framework that is more attuned to substantive issues in ethics research.

As artificial agents become more intelligent, they might choose methods and means that are not aligned with humans' and that could have catastrophic consequences for humans and for the rest of the planet (AI alignment problem).

The general aim of this chapter is to contribute to a human compatible solution the problem of AI alignment. The *human compatible* approach suggests that intelligent agents be not required to maximize a simple given reward function attached to single goals, but to maximize the realization of human preferences, which are essentially uncertain. Our idea aims to provide novel methods for norm- and value-based reasoning in AI systems, i.e., methods that will enable AI systems to comply not only with precise rules, but also with ethical principles and values, which are notoriously imprecise and hard to pin down.

This project contributes to the goal of preventing AI control by putting humans at the center of intelligent machine design, and does so uniquely by pooling expertise in philosophy, logic, and computer science.

Embedding of human values in artificial general intelligence: Problems and some possible solutions

This paper aims to contribute to the embedding of human values in artificial general intelligence (AGI), both formally and substantially.

Artificial general intelligence is the hypothetical intelligence of a machine which can learn and reason to the full range of human abilities and potentially beyond. Many scientists believe that AGI will be reached within the end of the current century. Artificial intelligent agents are instructed to optimize a reward function, which is thought to describe precisely the combination of goals to be reached. Even if it were possible to judge and weigh correctly all the subgoals, as agents become more intelligent (even without considering singularity scenarios), they might choose methods and means that are not aligned with

humans' and that could have catastrophic consequences for humans and for the rest of the planet: this is the so-called *AI alignment problem* (cf e.g. Bostrom, 2014; Russell, 2019).

The *human compatible* approach (Russell, 2019) suggests that intelligent agents not be required to maximize a simple given reward function attached to single goals, but to maximize the realization of human preferences, which are essentially uncertain. To reach this latter goal, current research is looking into *inverse reinforcement learning* (IRL), as originally studied in e.g. Ng and Russell, 2000. Contrary to reinforcement learning, in IRL reward functions are *not given*, but *have to be learnt*: the agent observes the behavior of humans and has to learn their objectives, aims and values (i.e. infer their "reward function"), so that the AI, being uncertain about the "real" reward function, is expected to behave conservatively (e.g. to ask for confirmation before taking wild initiatives).

There are least *three* orders of problems with this framework.

Problem 1: all you can ever observe is behavior, and it is hard to analyze it into the underlying *beliefs and values*, which are directly unobservable but (ii) there is also unobserved behavior (i.e. very rare or impossible), crucial for *counterfactual or counterpossible reasoning* (i.e. counterfactuals with (necessarily) false antecedents), which is relevant both to reach higher intelligence (cf. Pearl, 2000) and to ascribe responsibility; however, one of the most popular tool to handle counterfactual reasoning in AI (the causal Bayesian networks of Pearl, 2000) is based on intensional logics, which cannot handle non-omniscient agents and cannot entertain counterpossible scenarios;

Problem 2: (iii) modeling often relies on behavior that is assumed to be *optimal* (coherent, full knowledge, etc.), or agents that are idealized, while it should be able to understand when humans make mistakes; (iv) at any given time, there are many reward functions for which an observed behavior is optimal (*degeneracy*);

Problem 3: (v) even if the preceding strategy is successful, at a certain point the AI will have learnt all there is to learn, and therefore behave as a *utility maximizer*, thus recreating all the problems of misalignment, because current approaches to normative uncertainty (cf. William MacAskill and Ord, 2020) assign a probability distribution to value theories (utilities), thus assuming that there is one true utility function. In other terms, the uncertainty is epistemic, not ontological, and is stationary (i.e. agents' preference do not change over time).

~~Research questions, hypotheses, objectives and expected re-sults~~

The general aim of the approach sketched here is to contribute to a human compatible solution the problem of AI alignment as limited to issues (i)–(v) above. We aim to provide novel methods for norm- and value-based reasoning in AI systems, i.e., methods that will enable AI systems to comply not only with precise rules, but also with ethical principles and values. The general approach I suggest to reach the goal of preventing AI control is by putting humans at the center of intelligent machine design, namely by exploring and further refining the Human Compatible approach. This results in (super)intelligent machines being dependent on us at a deep level (due to the constant learning needed to reconstruct our values), rather than viceversa, and in integrating (super)intelligent agents in our *Wertanshauung*.

The approach has the following more precise objectives:

Objective 1: Integrate realistic/fine-grained counterfactual reasoning into inverse reinforcement learning, thus addressing Problem 1 (issues (i)–(ii)) and Problem 2, issue (iii).

Objective 2: Develop an account where human values are essentially (ontologically) imprecise, thus addressing Problem 2 issue (iv) and Problem 3 (issue (v)).

Objective 3: Validate and test the developed framework, implementing corrections when needed.

The expected result is both a formal and a substantial contribution to the current effort to reach artificial general intelligence which integrates computer science and philosophical expertise.

The more precise expected results are the following:

W.r.t. Objective 1: specify a hyperintensional semantics for counterfactual reasoning. First, such an approach would be able to discriminate between logical and necessary equivalents, and it is therefore finer-grained than intensional approaches. Such a feature is necessary both if intelli- gent artificial agents have to understand humans' preferences, given that humans have cognitive limitations precluding logical omniscience, and to avoid the optimality problem.

In order to develop a fine-grained approach to counterfactual reasoning, the framework of choice is truthmaker semantics, which has been proposed by Kit Fine (cf. Fine, 2017), and

which I further refined to be applied to deontic reasoning, the logic of responsibility, and rea- sons (cf. Faroldi, 2019).

W.r.t. Objective 2: Develop a formal approach to uncertain values using imprecise value functions. Once we know which states there are, which are merely possible, and which are impossible and how to counterfactually reason with them, we need to be able to infer reward functions. We need to know how agents value states. Given the set of states Ω, a suitably generated structure **F** of its subsets, we define **M** as a (finite) family of partial indexed signed measures μ^{α}_{i}, $i \in I$, s.t.μ^{α}_{i} :**F**→R,whichgenerateanintervalofvalues,ratherthanasinglevalue,foreachagent $\alpha \in A$.

W.r.t. Objective 3: Validate and test the developed framework with benchmark cases (such as the trolley problem). Apply the fine-grained framework so that agents can estimate their own re- sponsibility and study the consequences of this proposal at the normative and application level, by introducing a notion of legal personality for artificial agents and attenuating the notion of strict (or product) liability for the creators. Two necessary elements of responsibility ascrip- tions, according to the mainstream stances (Talbert, 2019) are the possibility of doing otherwise and the value or disvalue of a course of conduct. These elements open the possibility for AGI to participate in humans legal and moral practices of responsibility attribution.

Case 1: Paperclip Scenario Consider the paperclip scenario, where an AGI is tasked with maximiz- ing paperclip production. Unaligned scenarios where the AGI destroys the rest of the universe to pro- duce paperclip, prevents a switch-off, engages in wireheading, resists changes to the reward function are already ”solved” by the human compatible approach: by maximizing human preferences, which are uncertain, all these problems would be avoided. However, human preferences and values are learnt by observation of human’s behavior. But there’s no behavior to observe when it comes to very impor- tant but catastrophic events, that never realized before: the destruction of the human race, or the provoked extinction of entire species. As applied to the paperclip problem, the current project steps in to provide tools to (i) take into account reasoning with unobserved situations via counterfactual reasoning (what would happen if all iron in the universe would be consumed to produce paperclips? would it still make sense to maximize paperclip production if no humans lived anymore?), in (ii) a highly fine-grained (hyperintensional) way, by taking into account a range of values, rather than one, to take into account the problem of assigning values to unobserved states (the extinction of a

random species vs the extinction of an iron-eating bacterium if iron is depleted vs the extinction of a parasite species), to the fact that there are very different humans who might disagree, normative uncertainty, etc.

Case 2: Trolley Problem A famous benchmark case in moral philosophy is the trolley problem. There's a trolley about to hit *n* people, but you press a switch, the trolley modifies its direction and you only hit 1 person. Different moral theories have different positions on what to do.

How should an AGI behave? And how will it? In a traditional IRL setting, when confronted with a trolley problem, an autonomous vehicle should observe the behavior of experts and extract their reward function. We aim to test the framework developed in the first part of the proposed account using data collected by the Moral Machine experiment (Awad et al., 2020). A variant requires you to push a fat man in front of the trolley instead of pushing the switch button. Different moral theories have different positions on what to do: deontology-inspired theories recommend doing nothing and killing X people, consequentialist-inspired theories recommend switching track if $1 < X$.

In the setting developed in the current project, an autonomous vehicle would have to consider counterpossible scenarios, take into account their value. While this cannot be obviously observed, it is a highly relevant point that is used to compare different version of utilitarianism. Moreover, it will only consider value intervals, rather than single outputs, to off-set the chance that different moral theories might be right, or that an average of moral theories is taken while they are incomparable.

Methodology

The present sketched account intends to blend research in logic and computer science with research in informal areas of ethics and philosophy. This mix has the particular aim, on one hand, to dispense with simplifying assumptions that are quite widespread in technical fields (omniscience, that all behavior is observable, that there exists one true value function) by taking on board the complexity of humans' normative life and, on the other hand, to try and make more precise (and usable) some of theories and concepts of more humanistic disciplines (normative uncertainty, responsibility).

Instead of the traditional intensional approach based on possible worlds, which are consistent, complete, and flat, truthmakers are states that are not necessarily assumed to be consistent, complete and that are equipped with a mereological structure. This set-up is

more adequate for agents with limited capabilities having to reason with real-world information, which is possibly incomplete and contradictory, but can be expanded. Moreover, this framework has been shown to be able to handle well non-causal explanations (cf. e.g. grounding), which makes it an ideal choice not just for counterfactuals, but also for counterpossibles. The framework developed by Fine, 2012 will be integrated with the Bayesian networks approach developed in e.g. Pearl, 2000 to reach Objective 1, rather than with standard intensional semantics.

This project assumes that human values are essentially imprecise. This can capture the sort of value uncertainty that is thought to be necessary for alignment within the human compatible framework. In contrast to some of the current approaches to normative uncertainty (cf. William MacAskill and Ord, 2020), which assign a probability distribution to value theories (utilities), here each value function is imprecise, i.e. does not result in sharp values but in an interval of values.

Imprecise values, coupled with the ability to reason about what could have been, make it possible to know what should have been and what could and should be, all essential elements in the ascription of responsibility. Ascribing (and self-ascribing) responsibility to intelligent agents is one way to inte- grate in an ethical way humans and general intelligent agents, for it avoids the the problem of an agent turning into an utility maximizer: there is always the possibility of being responsible looming ahead.

The main conceptual limitation of the testing phase is that the framework, when fully developed, is about AGI, which we don't have yet. The testing will then have to be on smaller available compo- nents.

References

Awad, Edmond, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-Franç ois Bonnefon 2020 "Universals and variations in moral decisions made in 42 countries by 70,000 partici-

pants", *Proceedings of the National Academy of Sciences*, 117, 5, pp. 2332-2337. 2014

Bostrom, Nick

Superintelligence: paths, dangers, strategies, Oxford University Press.

Faroldi, Federico L. G.

2019 *Hyperintensionality and Normativity*, Springer, Dordrecht.

Fine, Kit

2012 "Counterfactuals without Possible Worlds", *Journal of Philosophy*, 109, 3, pp. 221-246.

Ng, A. Y. and Stuart Russell
2000 "Algorithms for inverse reinforcement learning," in *Proc. 17th Intl. Conf. on Machine*

Learning (ICML-2000), ed. by P. Langley, Morgan Kaufmann, pp. 663-670. 2000

Pearl, Judea

Causality: Models, Reasoning, and Inference, Cambridge University Press.

Fine, Kit

2017 "Truthmaker Semantics", in *A Companion to the Philosophy of Language*, ed. by Bob Hale, Crispin Wright, and Alexander Miller, 2nd ed., ms, Blackwell, London, pp. 556- 77.

Russell, Stuart

2019 *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking.

Talbert, Matthew

2019 "Moral Responsibility", in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Winter 2019, Metaphysics Research Lab, Stanford University.

William MacAskill, Krister Bykvist and Toby Ord
2020 *Moral Uncertainty*, Oxford University Press.

The Normative Risk Approach

This chapter discusses three notions of normative risk: (i) normative risk as a probable normative

or moral harm, (ii) normative risk as normative underdeterminacy, and (iii) normative

risk as norm-related existential risk. The report considers briefly an application of the

concept to artificial intelligence regulation, and, after a brief interlude on short- and

long-term strategies, introduces a few practical recommendations to tackle normative

risk.

Risk

The Stanford Encyclopedia of Philosophy (Hansson, 2018) distinguishes five main uses

of the 'word' risk, from unwanted event that might or might not occur, to the cause

of such an event, the probability of such an event, the statistical expectation value, to

the more widespread technical use of risk to refer to the fact that a decision is made

under conditions of known probabilities (as opposed to a decision that is made under

uncertainty).

If the technical definition is adopted, a lot will depend on how probability is conceptualized, and to what extent it is even possible to quantify probability for complex

interactive systems.

Standard decision theory has two main ingredients that interact: possible states and

values. The states are associated to uncertainty, which is represented with probability,

and values are represented trough utility. More recently, theorists have started wondering on how to represent uncertainty about values.

While the probability of states is compatible with objective and subjective interpretations of probability, this kind of normative uncertainty seems largely to be understood subjectively, e.g. epistemically.

But what about the case where the uncertainty is is objective, i.e. values are (metaphysically) underdetermined? Are there substantive differences from the case where values are just epistemically uncertain? Do we have to use a different formal framework? Enters normative risk.

Normative Risk

I distinguish between three notions of normative risk: (i) normative risk as a probable normative or moral harm, (ii) normative risk as normative underdeterminacy, and (iii) normative risk as norm-related existential risk.

Normative risk as a probable normative or moral harm Normative risk as a prob- 3 going to depend on the background normative (moral) theory. This notion seems to be a subset of a non-technical notion of risk as probable harm (unwanted event), where the harm in question happens to be normatively or morally relevant.

Normative risk as normative indeterminacy Normative risk as normative indeterminacy is further split in two. The former notion consists in value gaps, or situations that are normatively indifferent (lacunae) — which we refer to as normative underterminacy. The latter notion consists in value gluts, or situations that are normatively

qualified in incompatible ways, such a situation being both obligatory and forbidden (antinomies) — which we refer to as normative overdeterminacy. It is important to

note at this point that this indeterminacy is at the ontological level, not at the epistemological level, as is the case in the normative (moral) uncertainty literature.

Normative risk as norm-related existential risk Existential risk is risk that greatly imperils life on planet Earth as we know it, e.g. by leading humans to extinction or to the collapse of civilization via catastrophic events.

When it comes to norm-related existential risk, there are at least two ways to spell out the meaning of 'related': causation and omission. There are thus at least two notions of normative risk in this second sense: when norms and values increase the probability of an existential risk, or when they fail to decrease the probability of an existential risk. Studying normative risk consists in analysing how existing or proposed norms engender or fail to prevent future catastrophic events or existential risks. A prime example is the proposed EU AI regulations that fail to prevent developers from developing future general intelligent systems that result in a singularity and might be misaligned.

Further theoretical notes on normative risk The technical report in preparation considers two questions:

First, are the three notions of normative risk connected, and if they are, how?

Second, how to model them?

Preliminary, one should think that the former question has a theoretical priority, since if the two notions are connected, presumably the modeling should reflect this fact. However, pragmatically, I find it more useful to start from the less familiar notion, i.e. that of normative risk as norm-related existential risk.

Strategies to Mitigate Normative Risk

Mitigation of normative risk is a complex process. It starts with the identification of normative risk, followed by analysis, quantification, and finally the elaboration of mitigating strategies. The implementation, instead, occurs through usual means. We distinguish between long term and short term mitigation strategies.

4.1 Long-Term

Ord, 2020 and some colleagues at the Future of Humanity Institute have put forward a simple but effective analysis to decide which (existential) risk to tackle first. There are three criteria: Importance, Tractability and Neglect.

Importance is how much a specific risk contributes to the overall risk — what is the value of mitigating it?

Tractability is how easy it is to mitigate or solve the problem.

Neglect is a measure of how much attention and resources are devoted to solve it. Keeping one or more factor fixed, even if it is hard to give precise estimate of any, helps with determining a strategy to deal with risks in the long term.

There are no reasons to believe these criteria should not work for the first and third notion of normative risk.

4.2 Short-Term

The same factors are used to establish short term strategies when dealing with a specific normative risk, but it will refer, this time, to different strategies to solve or mitigate it.

4.3 An example: AI and Risk Management in the proposed EU AI Act

The EU has strong regulatory force. Not all its regulations are not immediately valid

in its member states, but they tend to be reproduced in other parts of the world, and major companies spontaneously comply as the EU is the largest world market (according to some metrics). This was the case with the EU regulation on privacy, the GDPR. The new proposed regulation on AI, therefore, holds the promise of holding the same or greater sway than the GDPR, both at the public (e.g. non-EU states adopting something very similar) and the private level (e.g. companies which don't have to but voluntarily comply).

Contrary to many proposals leading up to it in the preceding years, the proposed EU AI Act does not take the route of assigning (some degree of) legal personhood to artificial agents, on which some degree of responsibility can be based. Instead, the whole act rests on a notion of risk management.

To be any effective, the EU AI Act has to define very precisely what risk is, what is required to prevent it, and how to repair eventual damages.

The Act defines four levels of risk:

1. Unacceptable risk:forbiddenactivities
2. high risk: obligatory requirements and pre-conformitycheck
3. low risk: transparencyrequirements
4. minimal risk: transparency requirements.

Unfortunately, contrary to the GDPR, the proposed EU AI Act does not engage in providing a general definition of risk, nor a general strategy to identify risks. Instead, risky systems are determined with lists, which include sensible areas of human society

and life.

This is itself risky: first, because many activities which are not risk-prone will be classified as risky and perhaps banned, with detriment to advancement; second, many activities that are not in the list but are risky, will be generally allowed. This is an extrinsic characterization of risk, whereas an intrinsic one is needed for efficacy and generality.

The proposed EU AI Act, therefore, exemplifies at least two notions of normative risk: its adoption (in the current form) would increase the probability of moral harm, thus exemplifying the first notion of normative risk; and its adoption (in the current form) would also increase the probability of existential risk (i.e. AGI taking control) or fail to decrease it,

thus exemplifying the third notion of normative risk.

This is a good point to mention that while the first and third notion seem to coincide in this case, they don't: in fact there may be moral harms that are not existential risks; and existential risks that are not morally relevant, at least according to some moral theories.

But let's focus a bit more on the risks of the EU AI Act and how the normative risk approach can help.

Let's consider two starting points:
First, the possibility of *singularity*: machines that, having reached the stage of artificial general intelligence (AGI), or superintelligence, would give rise to an "explosive" moment that could radically and irreversibly change the planet and human civilization .

Second, the so-called *control problem*: how can we make sure that general intelligent, or so-called superintelligent, agents do not take control over us and the world?

This is an existential risk, that is, an event that increases the probability of eradicating biological intelligent life on the planet, or a similar catastrophic collapse of civilization.

The proposed EU AI Act fails to take into account general intelligent systems: first, because it fails to take into account AI systems that do not have a concrete intended application, second, because it only takes into account objectives specified by humans, therefore ignoring the problem of agents that will set their own objectives or instrumental sub-objectives.

Therefore, (i) it fails to properly take into account the control problem due to potential general AI; (ii) even more modestly, fails to account for existing non-specific systems, like GPT-3.[79]

So the proposed regulation(s) will run after and follow, rather than shape, technological development

But alignment has to be built in before a singularity is reached, i.e. before it is too late for humans to gain back control.[80]

What to do?

According to my general theory, that I call the Normative Risk Approach, one has to individuate and mitigate those events that increase (or fail to decrease) the probability of an existential risk, events that are tied to regulation, or absence thereof.

In this concrete case, we have to already consider now the possibility of a general intelligence that might be misaligned, and not just build transparency requirements or human oversight, but already build incentives and sanctions in an agent's reward system in a way not dissimilar from the criminal law.

[79] the Commission does not foresee AGI systems is also revealed by the provisions on human oversight

- Take control and prevent itself to be switched off
- Manipulate the human controllers

[80] it makes sense to legislate already now to safeguard and prevent this possibility, even if remote, rather than complain later, when too late.

5 Practical Recommendations

Normative Risk Lab

A Normative Risk Lab would perform theoretical research on defining and modeling normative risk, elaborate general mitigation strategies, and potentially elaborate policy 8 recommendations.

5.2 Normative Risk Consulting

There are two addressees: private companies and governments. The former would benefit from a compliance check on existing legislation, the latter would benefit from a preliminary check on proposed legislation.

Normative Risk Independent Administrative Authority

Several countries have independent authorities (i.e. fairly independent from the government, judiciary, and parliament), with regulative, inspective, and certificatory powers, in fields ranging from insurance, freedom of communication, stock market, energy, pension funds, corruption and prisoners. One idea is to have a Normative Risk Independent Administrative Authority with similar powers, that preemptively analyses proposed legislation for normative risks, publishes binding guidelines for public administration, and can do posterior checks.

The pros: such an institution will have real powers to intervene on public and private initiatives, both before and after. The cons: while being independent from governments, such an authority will still be subject to legislative authority, and therefore

political pressure, thus potentially compromising its effectiveness.

deablehormativeormoralharmisperhapsthemostbasichotion. Itsexactdefinitionis

References

Bykvist, K.

2017 "Moral Uncertainty", *Philosophy Compass*, 12, 3, e12408, doi: 10.111 1/phc3.12408.

Dietrich, F. and B. Jabarian

2020 "Expected value under normative uncertainty", ms.

2022 "Decision Under Normative Uncertainty", *Economics and Philosophy*.

Faroldi, F. L. G.

n.d. "Modeling Value Disagreement via Imprecise Measures", 2017.

2021a "Towards a Logic of Value and Disagreement via Imprecise Measures",

Bulletin of the Section of Logic, 50, 2, pp. 131-149, doi: 10.18778/0138- 0680.2021.07.

2021b "General AI and Transparency in the EU AI Act", *i-lex*.

Hansson, S. O.

2018 "Risk", in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, Fall 2018, Metaphysics Research Lab, Stanford University.

MacAskill, W.

2016 "Normative Uncertainty as a Voting Problem", *Mind*, 125, 500, pp. 967- 1004.

Ord, T.

2020 *The Precipice. Existential Risk and the Future of Humanity*, Blooms- bury, London.

William MacAskill, K. B. and T. Ord

2020 *Moral Uncertainty*, Oxford University Press.

Glossary

Weak AI: programs that do not experience consciousness or do not have a mind in the same sense people do, but can (only) act like it thinks and has a mind and consciousness

Strong AI: ability of an intelligent agent to understand, feel or think like a human; sometimes is thought to require consciousness; "computers given the right programs can be literally said to understand and have other cognitive states. (Searle 1980: 417)"

Narrow AI: ability of an intelligent agent to learn and perform a specific task, often with at least human proficiency

General AI: ability of an intelligent agent to learn and perform any intellectual task that a human being can

Superintelligence: a hypothetical agent that would possess intelligence far surpassing that of the brightest and most gifted human mind.

singularity: a hypothetical point in time at which technological growth becomes uncontrollable and irreversible, resulting in unforeseeable changes to human civilization

misalignment (minimalist): an AI A to be misaligned with a human H if H would want A not to do what A is trying to do (if H were aware of A's intentions)

alignment (maximalist): an AI that incorporates values and behaves morally (or legally).

Control problem: we make sure that general intelligent, or so-called superintelligent, agents do not take control over us and the world. This would be an existential risk.

Existential risk: an event that increases the probability of eradicating biological intelligent life on the planet, or a similar catastrophic collapse of civilization

Normative risk (regulatory sense): events that increase (or fail to decrease) the probability of an existential risk, events that are tied to regulation, or absence thereof.

Artificial General Intelligence:

The idea of singularity is that if the trajectory of artificial intelligence reaches up to systems that have a human level of intelligence, then these systems would themselves

have the ability to develop AI systems that surpass the human level of intelligence, i.e., they are superintelligent (see below). Such superintelligent AI systems would quickly self-improve or develop even more intelligent systems. This sharp turn of events after reaching superintelligent AI is the singularity from which the development of AI is out of human control and hard to predict (Kurzweil 2005: 487).

Bibliography

ALPA, G. (2020) a cura di, *Diritto e intelligenza artificiale*, Pacini, Pisa.

ARTEMOV, Sergei and Melvin FITTING, "Justification Logic", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. ZALTA (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/logic-justification/>

B. GOODMAN and S. FLAXMAN, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2016) ICML Workshop on Human Interpretability in Machine Learning, arXiv:1606.08813 (v3); (2017) 38 AI Magazine 50.

BABIC, B., I. Glenn COHEN, Theodoros EVGENIOU and Sara GERKE, The Case Against Explainable Medical Machine Learning, *Science*, 2021.

BEN-DAVID, S., HRUBEŠ, P., MORAN, S. *et al.* Learnability can be undecidable. *Nat Mach Intell* 1, 44–48 (2019). https://doi.org/10.1038/s42256-018-0002-3.

Benjamin SCHNIEDER, A logic for 'because', *Review of Symbolic Logic* 4 (3):445-465 (2011).

Billi, M., A Symbolic Approach For Ensuring Fairness in AI. *i-lex* 14,1 (2021).

Burrell, J. 'How the machine 'thinks': Understanding opacity in machine-learning algorithms' (2016), *Big Data and Society* DOI: 10.1177/2053951715622512

BURRELL, J. 'How the machine 'thinks': Understanding opacity in machine-learning algorithms' (2016) *Big Data and Society* DOI: 10.1177/2053951715622512

Christian, B., *The Alignment Problem*, Norton, 2020.

CHRISTIAN, Brian, *The Alignment Problem*, Norton, 2020.

DORIGO, S. a cura di, *Il ragionamento giuridico nell'era dell'intelligenza artificiale*, Pacini, 2020.

Erika PUIUTTA, Eric MSP VEITH, Explainable Reinforcement Learning: A Survey. https://arxiv.org/abs/2005.06247

Faroldi, F.L.G (2021a), Considerazioni filosofiche sullo statuto normativo di agenti artificiali superintelligenti, Revista Iustitia, 9, 2021.

Faroldi, F.L.G (2021b), Trasparenza dell'algoritmo e deep learning. Note logiche a margine della proposta di regolamento sull'Intelligenza Artificiale (Artificial Intelligence Act) della Commissione Europea e di un'ordinanza della Corte di Cassazione, Revista Iustitia, 10, 2021.

FAROLDI, F.L.G, *Partial Reasons*, ms, 2021.

Faroldi, F.L.G, *Responsabilità e ragione.* Satura editore, 2020.

FAROLDI, F.L.G., *Hyperintensionality and Normativity*, Springer 2019.

Good, I. J. ,1965, "Speculations Concerning the First Ultraintelligent Machine", in *Advances in Computers*, vol 6, Franz L. Alt and Morris Rubinoff, eds, pp31-88, 1965, Academic Press.

Goodman, B. and Flaxman, S., 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2016) ICML Workshop on Human Interpretability in Machine Learning, arXiv:1606.08813 (v3); (2017) 38 AI Magazine 50.

Alexandre HEUILLET, Fabien COUTHOUIS, Natalia DÍAZ-RODRÍGUEZ, Explainability in deep reinforcement learning, *in* Knowledge-Based Systems, December 2020.

Juozapaitis, Z. et al., Explainable reinforcement learning via reward decomposition. URL: http://web.engr.oregonstate.edu/~afern/papers/reward_decomposition__workshop_final.pdf.

Krishnan, M., Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning, *Philosophy & Technology*, 33:487–502, 2020.

Lipton, Z. C., "The Mythos of Model Interpretability," https://arxiv.org/pdf/1606.03490.pdf, Jun. 2016.

Maya KRISHNAN, Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning, *Philosophy & Technology*, 33:487–502, 2020.

Ngo, R., "AGI Safety from first principles", ms, 2020.

Numerico, T., *Big data e algoritmi.* Carocci, Roma, 2021.

NUMERICO, Teresa, *Big data e algoritmi.* Carocci, Roma, 2021.

Ord, T., *The Precipice*. Bloomsbury, 2020.

Russell, S. and Norvig, P., Artificial Intelligence: A Modern Approach, 4th Edition, Pearson, 2020.

Russell, S., *Human Compatible*, Viking, 2019.

S. WACHTER, B. MITTELSTADT, and L. FLORIDI, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 IDPL 76.

Sara GERKE, Timo MINSSEN and I. Glenn COHEN, Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare. In Adam BOHR, Kaveh MEMARZADEH (eds.), *Artificial Intelligence in Healthcare*, Elsevier, 2020.

Stuart RUSSELL and Peter NORVIG, Artificial Intelligence: A Modern Approach, 4th Edition, Pearson, 2020.

Stuart RUSSELL, *Human Compatible*, Viking, 2019.

Swapnil Nitin Shah, Addressing the interpretability problem for deep learning using many valued quantum logic, https://arxiv.org/pdf/2007.01819v1.pdf

Swapnil Nitin SHAH, Addressing the interpretability problem for deep learning using many valued quantum logic, https://arxiv.org/pdf/2007.01819v1.pdf

TADDEI ELMI, Giancarlo, e CONTALDO, Alfonso (a cura di)., *Intelligenza Artificiale. Algoritmi giuridici.* Ius condendum *o "fantadiritto"?* Pacini Giuridica, Pisa, 2020.

Z. JUOZAPAITIS, A. KOUL, A. FERN, M. ERWIG, F. DOSHI-VELEZ, Explainable reinforcement learning via reward decomposition. URL: http://web.engr.oregonstate.edu/~afern/papers/reward_decomposition__workshop_final.pdf

Zachary C. LIPTON, "The Mythos of Model Interpretability," https://arxiv.org/pdf/1606.03490.pdf, Jun. 2016.