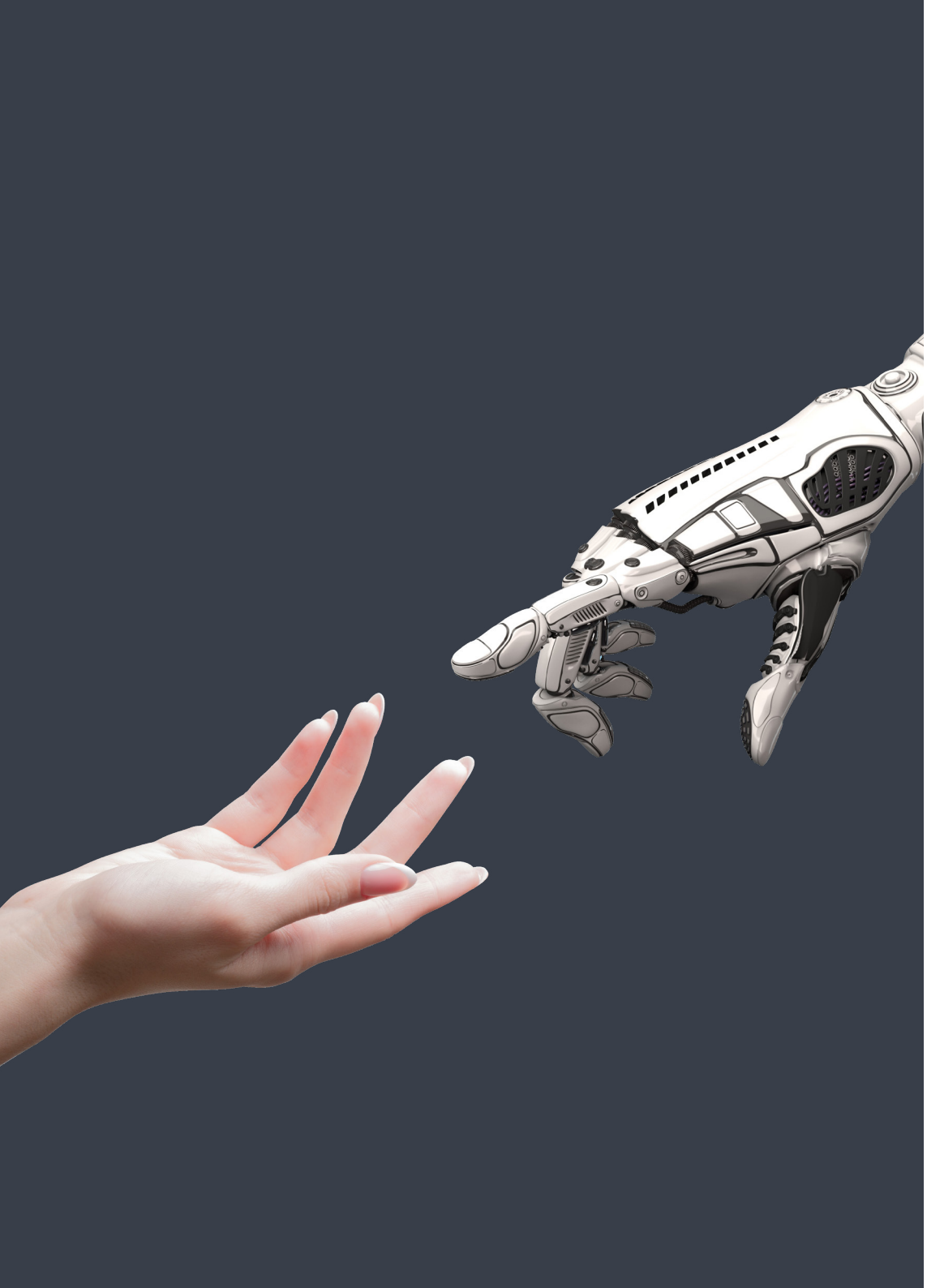


Artificial Intelligence

An exploration around morals and ethics in the development of AI

EARS

European Academy on Religion and Society

About this whitepaper

The development of Artificial Intelligence (AI) concerns all domains, ranging from politics to religion, and from ethics to philosophy. During our round table discussion on the topic, we explored how morals and ethics are taken into account in, for example, self-learning systems and humanoids. In addition, we discussed expected future developments in the field of AI.

This whitepaper is based on the outcomes of our virtual round table meeting. Several articles were written as an inspiration for the round table discussion. We would like to acknowledge the authors of these articles, the round table report, and the introduction to this whitepaper: R. Anthony Buck, Ghila Amati, Muhammad Faisal Khalil, Thom Lamers, and Prof. dr. Matthias Smalbrugge

EARS

Table of Contents

EARS

Artificial Intelligence:

Introduction written by Prof. dr. Matthias Smalbrugge

EARS

In the beginning God created mankind, but men created a better version of mankind: AI. AI is indeed a trending topic, it is never far away and it is highly relevant. Yet, it is also a topic that may vary depending on the definition one adopts. Do we speak about the increasing influence Big Tech is gaining on us? About independent drones? About dancing robots and thus about the vanishing border between humans and humanoids? All these angles are possible approaches and they can all be considered to belong to AI. Indeed, AI offers fascinating perspectives in different domains. Moreover, it is increasingly present in our daily lives. Suffice it to think of the highly personalised advertisements we all receive.

This omnipresent quality of AI reveals in any case that 'intelligence' can never be conceived of without a context. It is not an isolated experiment in a laboratory, it is situated in a certain context and it is this context what matters. There is a social context, a political one, a moral one, and so on. Such contexts can stimulate the development of AI, but they can also be an obstacle or raise questions. One can be very enthusiastic about the use of AI in the field of e.g. medicine, but one can also have moral doubts regarding the influence of Big Tech on democratic structures.

These contexts also reveal we can no longer consider AI as a purely technical revolution, belonging only to certain disciplines. We have to think of a development that pervades all disciplines, including humanities. Meaning that the so-called digital humanities (DH) no longer function as a subsidiary discipline, to be used just in order to create e.g. powerful databases, but that DH allow us to discover new fields of research within the domain of humanities.

Context and growing influence, that is what determines the debates on AI. We are confronted with questions only few people could have foreseen. The most pressing ones, however, seem to belong to the domains of ethics and politics, as well as to the field of philosophical matters. As a network of European Faculties of Theology and Religious Studies, we consider these developments also as pertaining to the domain of religion. What we therefore want to discuss are the complicated issues laying ahead of us. Seemingly, the boundaries between traditionally separated domains are vanishing. Suffice it to think of the influence of Big Tech on democracy, suffice it to think of the distinction between human and non-human, to quote just two examples of disappearing boundaries. In technical terms, ontological distinctions seem to change and morality and transparency are difficult to apply.

Artificial Intelligence:

Report written by Thom Lamers

EARS

Artificial Intelligence (AI) and religion are two subjects that are not often connected. Religion has been around for quite a while, while AI is yet to reach its great heights, according to some. Both, however, are very present in our current world. In all kinds of contexts, we are able to observe a shimmer of that connection. For instance, when we talk about morality, about responsibility, or about freedom. With the ever-advancing area of AI, we can be certain that this phenomenon will leave its mark on other fields as well. That poses some interesting questions. Some of the most pressing ones to be ethical, political, and philosophical. For instance, how will morals be instilled in self-learning systems and humanoids? What is the role of Big Tech in (future) developments? And how do the Roman-Catholic criteria concerning AI play a role? We discussed these topics and issues with participants from various religious backgrounds.

To have or not to have: morality

Morality has always played an important role in the development of humanity. It is shaped through feedback from the environment and influences the way we make decisions. The language around AI shapes our hopes and fears. It determines whether it is something to look forward to, or something to fear. When discussing morality and AI, it is useful to distinguish between the current and the future state of AI. Currently, it is inconceivable that any sense of morality would be assigned to AI. This is most certainly the case when the debate is about life or death. However, currently unimaginable things will perhaps become reality in the future. In the past, humans have done marvelous things without AI. Could machines potentially develop a higher moral than humans, since they have access to instant information and are able to decide without the complications of the human mind? Would we ourselves perhaps learn more about ourselves when working with machines, like looking in a highly intelligent mirror?

The question remains, however, whether it is even possible to talk about morality in AI. Some would argue moral actions to only be possible when there is a certain degree of freedom. As humans program decision-making systems in machines, one could argue that the notion of freedom is therefore hard to find. And if AI, by default, is not free, how would we make it responsible for its actions? Others would argue that we have not even mapped out our own behavior yet. For instance, we have not yet agreed on whether our behavior is deterministic or whether a degree of freedom is involved. The question arises if assigning morality to AI would not first require us to determine our own behavior. In any case, the topic of morality deserves some thoughtful consideration

Learning about freedom

Morality precedes the notion of choice. Are we able to make our own decisions? Are humans ever absolutely free? It is arguable that our decisions are the product of our biological makeup (hardware as you will) and our experiences through life (our software). What does that say about us? Perhaps we experience a sense of freedom as a result of what we learned in the past. When machines have the same possibility, minding the anthropomorphic trap, will the similarity between mankind and machines be even closer than we think?

What we do know is that the way in which AI learns is different from human learning. Humans learn from experiences, but we can only have those experiences in set times. AI can learn in completely different structured and complex ways. Machine intelligence will be different from human intelligence. When talking about ethical AI, we need to discuss moral learning, especially with today's knowledge of possible discrimination and bias in designing systems. How can we make machines learn responsibly? This brings us back to the notion of morality: neither innate nor divine, but something we acquire all the time during our lives. And therefore perhaps achievable for machines?

A moral imperative: empathy

A notion in sufism: 'God created man because man knows himself in the eyes of others'. Following this notion, AI could possibly provide us with a new kind of awareness and self-reflection. In which case we could ask: 'what do you think of me?' This question stresses another important factor: emotional intelligence. Our ability to feel, connect to, and understand each other. To invoke a theory of mind. If showing emotions means understanding the reasons for having emotions, how could we model that in a dual-system AI?

Currently, and in the near future, AI will probably not be capable of independent empathy but is able to draw on data that replicates human empathy. For example in California, where lonely people can interact with robots who read facial expressions. However, we do have to ask ourselves the higher moral question of whether we want to support robotic companionship for (lonely) people. Underneath might lay a possible shallowness that needs to be included in discussions on the matter. This calls for a moral imperative on creators and scientists to be honest and open.

Maintaining democratic control over technology

AI has the potential to connect us, but it also has the power to divide us. A current flaw of the internet is that it is a one-stop shop. This creates silos in which people think similarly about what the world looks like. Combine the notion of groupthink, and we close ourselves off within these bubbles. After all, there is no point in being on a seperate platform if nobody else is there. This is the consequence of a monopolistic situation where

“[C]urrently unimaginable things will perhaps become reality in the future.”

some people have massive followings on one platform and disappear on another medium.

A possibility of AI is to help us connect and somehow accomplish getting people out of their bubbles and operate somewhere else. This is like protecting people from only similar input and purposefully giving them other perspectives. Freedom of information is only given if connections between different pieces of information can be made. The EU is encouraging research that allows multi-stop shops; an ecosystem of interconnected search engines, so that we are no longer hostage to the value that companies embed in their algorithms. This would decrease our social dilemma and increase our connectedness.

Principles of the church

Despite differences, an open dialogue between interdisciplinary groups can provide new perspectives. Each of the principles that the Roman Catholic Church has written down (transparency, inclusion, responsibility, impartiality, reliability, security and privacy) can be discussed elaborately, but the common narrative is striving for the best version of AI. This is where it becomes interesting. Concerning the principle of privacy, for example, can AI be efficient without knowing things? Can it offer us the best music without knowing us? And concerning the principle of transparency, would people be able to game the system if it becomes too transparent? And what is actual transparency? Is it a company that uses open source algorithms, or is it a company that simply states they are transparent? All things considered, these are some serious questions concerning an ideal, and a real situation.

The principles mentioned by the church are not specific to the Roman Catholic Church. They are already widely embraced by the AI community. A valid remark, however, is that trust is not named within these principles. Trust as a basic principle for the relation between humans and AI. In religious context, trust is already mentioned in different ways. The Koran, for instance, mentions that ‘men were presented with something of a trust’. God is self-imposing the idea of ‘I am not going to control men’, to the extent that men can even spit on God. Perhaps this is an example of how we as creators could build a notion of trust in our creation.

Many students trust AI to be able to make a positive impact, especially in the possibilities of detecting exclusion and discrimination. However, the academic field is very different from the corporate field. They have different motives for further development. The question is whether people will long accept that corporate AI will keep certain secrets when it comes to their development. After all, you will probably not eat food that you are not familiar with or do not know the origin of.

Coming back to the principles and the issue of transparency, could there be a scenario where we would all have a ‘private robot priest’ which helps us read all the information that we ourselves cannot attain? Maybe we need another entity in the form of AI that will handle our information overload. Only time will tell.

The role of religion

There are broad trends in how established theology groups respond to AI. Some religious groups even see AI as another narrative of the creation of the world. In any case, religious practices have to be adjusted to the new environment. That is how it has always worked in the past. In terms of religion, however, the only being above human beings is God. So, perhaps we need a new commitment to spiritual sources such as the Bible or the Koran: one that states that human beings will not put AI above them, as a call for modesty.

Religion transforms itself. We are a cultural narrative that we call religion. We are no longer what we were. The same goes for other sciences. Perhaps some think that religion remains as it was. However, we are no longer this sacred domain facing the profane. We are one of the narratives influencing our society that is undergoing the transformation of religion.

Appendix

Church and science over the years

Written by Ghila Amati

EARS

The complicated relationship of Church and science throughout the centuries

The relationship between science and religion has never been easy. This article will mostly explore the relationship between science and the Roman Catholic Church (RCC) throughout the years.

The RCC has always had an ambivalent approach to science. On the one hand, the Church has been one of the main and most important supporters of science throughout the centuries. It has often supported science through financial donations or by founding schools, universities, and hospitals. In addition, many of the most remarkable scientists of the past have been clerical figures. Historian Pierre Duhem argues that Catholic mathematicians and philosophers such as John Buridan, Nicole Oresme, and Roger Bacon could be considered the pioneers of modern science.[1]

Officially, the RCC states that there is absolutely no contradiction between science and the Christian faith and between reason and faith. In the document Catechism of the Catholic Church published for the Catholic Church by Pope John Paul II in 1992, the following is stated:

> "Though faith is above reason, there can never be any real discrepancy between faith and reason. Since the same God who reveals mysteries and infuses faith has bestowed the light of reason on the human mind, God cannot deny himself, nor can truth ever contradict truth. ... Consequently, *methodical research in all branches of knowledge provided it is carried out in a truly scientific manner and does not override moral laws, can never conflict with the faith,* because the things of the world and the things of faith derive from the same God."[2]

It is important to note that the statement that there are no contradictions between the RCC and science could bring different and even opposed approaches to the new discoveries of science. A possible reaction of the Church could be the attempt to harmonise scientific discoveries and religion — proposing for instance allegorical interpretations of the scriptures so that scientific discoveries will not contradict the Bible. However, another possible reaction could be the negation of the truthfulness of the scientific discovery itself when it does not fit the dogmatic principle of faiths of the Church.

For this reason, on the other hand, on the side of this more positive view of the approach of the Church to science there is also the 'conflict thesis.'[3] This thesis states that there is a historical and fundamental conflict between the Catholic Church and science. The main example proposed by the conflict theorists is the one of the trial of Galileo. According to this view, Galileo is "a symbol of a very fundamental conflict — the conflict between science and religion, between reason and faith."[4] Let's now look more deeply into the Galileo affair.[5]

The Galileo Affair

Most of us are familiar with the Galileo affair. In 1610, Galileo published a book called *Sidereus Nuncius* (Starry Messenger). In the book, Galileo endorsed the heliocentric theory, i.e. the astronomical model according to which the Earth is not at the center. This theory was in opposition to the accepted theory of the time and argued that Earth and planets revolve around the sun, which is at the centre of the solar system. Galileo's discoveries were in opposition with the belief of the Church. At that time, the Church believed in the Aristotelian geocentric view of Earth that was in line with the literal interpretation of the Bible in many places. The geocentric view believed Earth was at the center of the universe and that all heavenly bodies revolve around the Earth.[6]

In 1616, the Catholic Church and the Inquisition declared heliocentrism to be "formally heretical." Books that supported the heliocentric theories were forbidden and the Church prohibited Galileo to teach, explain, and uphold heliocentric views.[7] In 1633, the Roman Inquisition put Galileo on trial and accused him of "vehemently suspect of heresy," sentencing him to indefinite imprisonment. Galileo lived under house arrest until his death in 1642.[8]

Less known is that the Galileo affair and the opposition to Galileo on the side of the Church continued for centuries. Although in 1758 the Catholic Church

took out the Index of Forbidden Books those books that advocated for heliocentrism, it still did not openly revoke the accusation of the Inquisition against Galileo in 1633.[9]

Another milestone of the Galileo affair was in 1979, when Pope John Paul II started a Pontifical Interdisciplinary Study Commission to study the case of Galileo. He said he hoped that "theologians, scholars and historians, animated by a spirit of sincere collaboration, will study the Galileo case more deeply and in loyal recognition of wrongs, from whatever side they come."[10] However, the commision did not arrive to a clear conclusion and therefore did not achieve the goal and hopes of the pope expressed in 1979.[11]

Lastly, in 1992, the newspaper *L'Osservatore Romano* reported that the Church had finally changed its mind on Galileo affair:

> "Thanks to his intuition as a brilliant physicist and by relying on different arguments, Galileo, who practically invented the experimental method, understood why only the sun could function as the centre of the world, as it was then known, that is to say, as a planetary system. The error of the theologians of the time, when they maintained the centrality of the Earth, was to think that our understanding of the physical world's structure was, in some way, imposed by the literal sense of Sacred Scripture."[12]

Catholic Church and evolution

Less problematic is the relationship of the Church with the theory of evolution. This theory found — on the basis of scientifically gathered evidence — that species evolve and survive by adjusting to their environment better than other (comparable) species. As a consequence of the theory, the Genesis creation story proved not to be a faithful account of the process of creation. Evolution, therefore, contradicted the creation story as it was explained in Genesis and undermined that there was a supernatural and intelligent entity that created the world, the different species and human beings all at once. Nevertheless, the Church never explicitly negated this theory.[13]

Pope Pius XII 1950, in the encyclical "Humani Generis" argued that a coexistence between Catholic teachings on creation and evolutionary theory is possible.[14] Moreover, in 1996, Pope John Paul II said that Darwin's theory of evolution could be "more than a hypothesis." Additionally, Cardinal Joseph Ratzinger (who later became Pope Benedict XVI) said in 2002 that the "converging evidence from many studies in the physical and biological sciences furnishes mounting support for some theory of evolution to account for the development and diversification of life on Earth."[15] Finally, current Pope Francis reaffirmed that according to the Roman Catholic Church, "evolution in nature is not inconsistent" with church teaching on creation.[16] [17]

“[T]he Vatican recognises that the development of AI technology is already too advanced to be halted completely”

Catholic Church and the AI challenge

The last issue to analyse is the relationship between the Church and Artificial Intelligence (AI) technology. AI refers to “systems that display intelligent behaviour by analysing their environment and taking actions — with some degree of autonomy — to achieve specific goals.”[18]

Since the technology of AI is relatively new, the approach of the RCC to AI is not well known. At the beginning of 2020, the Vatican released a document entitled *Rome Call for AI Ethics*. This document seeks to provide a direction for AI to benefit society, rather than undercut humanity. From this document it is possible to form a first idea of the position of the Vatican in the AI technology debate.[19]

It appears from the document that the Vatican recognises that the development of AI technology is already too advanced to be halted completely. Moreover, the Church does not negate the many possible advantages this technology could bring to our society at large. Nevertheless, the Church is also aware of the many dangers that are embedded in the development of this technology. It stresses the importance of being aware of these dangers ahead and wants to cope with them in the best way possible. The Church is mostly worried about *ethical* and *social* problems that may develop as a result of this technology and the document establishes six general principles which should be implemented in the production of AI technology. These principles are the principle of *transparency*, i.e. AI technology must be explainable to all; the principle of *inclusion*, that is “the needs of all human beings must be taken into consideration so that everyone can benefit, and all individuals can be offered the best possible conditions to express themselves and develop”;[20] *responsibility*, i.e. those who design and deploy AI must proceed with responsibility and transparency; *impartiality*, i.e. AI should be developed in a way that will not cause bias among different groups; and *reliability, security* and *privacy*, i.e. AI systems must be both reliable and safe and should not infringe on human privacy.[21]

How is the Church’s current reaction towards AI similar or different from the past?

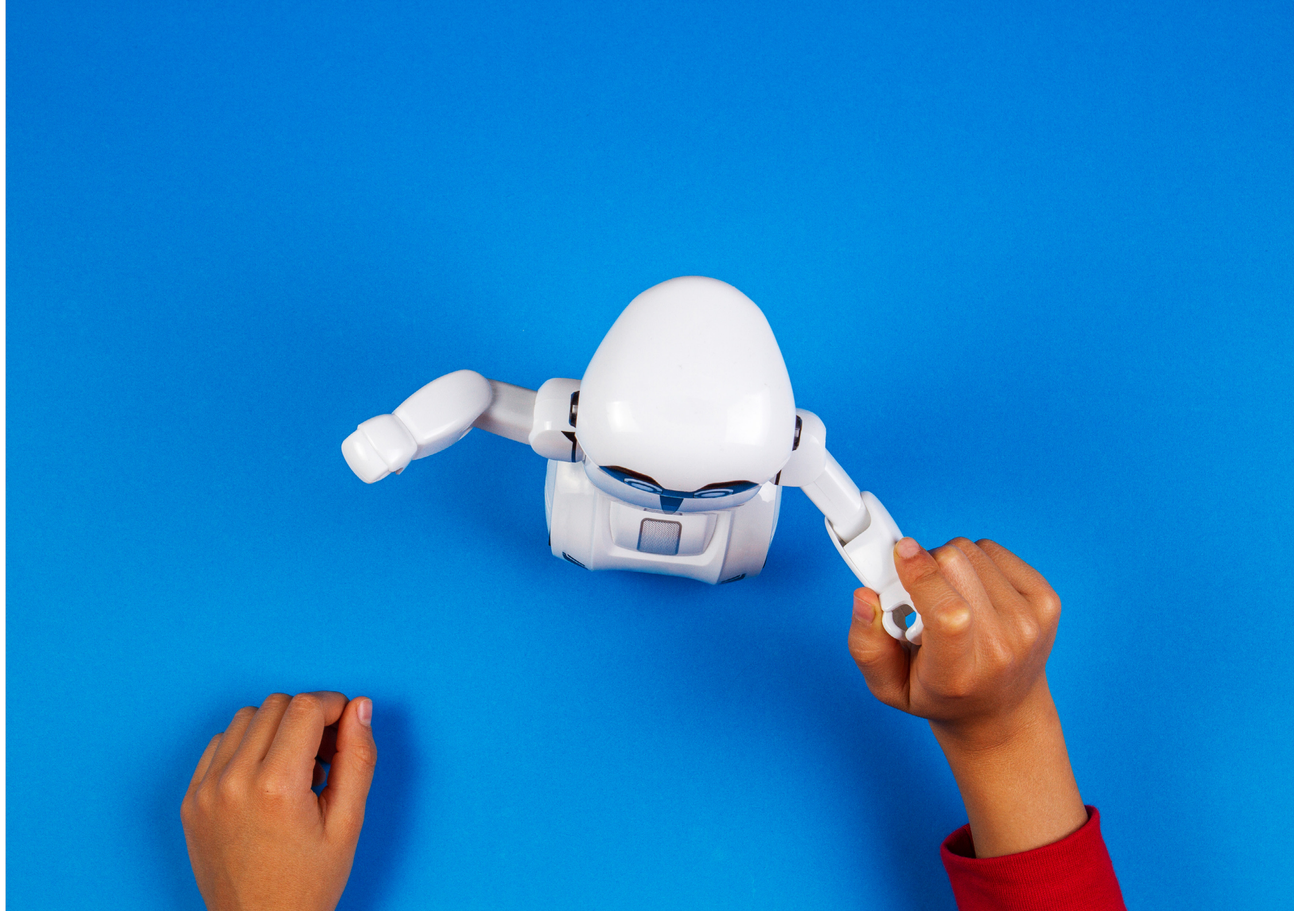
The reaction of the RCC towards the development of AI appears to be different from the approach the Church showed towards scientific discoveries in the two examples brought above. This time the Church does not seem interested — at least right now — in the way AI technology contradicts some theological fundamentals of the Christian Faith (as in the case of Galileo). Neither has it tried to show that AI technology developments and the beliefs of the Church can be harmonised (as in the case of the theory of evolution). The reason for this may be that AI technology is already spreading everywhere and the RCC is making use of it itself. AI therefore forces the RCC into a different position than that of harmonisation or persecution (the RCC does not have that power anymore anyway), namely cooperation and critical reflection. The RCC could grow into a new role here: as a critical dialogue partner. Thus, while in the past the Church was worried about the possible impact of new scientific discoveries over the beliefs of the faithful and over the theological principles and dogmas of its faith; this time, the worries of the Church are more about *social* and *ethical* implications of this technology over society at large.

The church is not yet focusing on the theological and dogmatic challenges AI may pose to the Christian faith. Some of these issues will be analysed in the article *Can AI Replicate Religious Leaders and Rituals?* We will see that AI in fact raises the theological issues of free will, repentance, and the afterworld. This means that soon enough the Church will need to develop a critical position over these matters as well.

Can you teach new robots old tricks?

Written by R. Anthony Buck

EARS



The year is 2035. The city is Chicago. The inventor of advanced AI robots has been found dead. An advanced AI robot named Sonny is Detective Spooner's prime suspect.

Spooner: The other day at the station, you said you had dreams. What is it you dreamed?

Sonny: I see you remain suspicious of me, detective.

Spooner: Oh, well, you know what they say about old dogs…

Sonny: No, not really.

This is a key scene in the 2004 blockbuster *I, Robot* directed by Alex Proyas.[1] It raises a lot of questions about AI and ethics. Can you make sure AI knows and does what is ethical? In the scene, Detective Spooner alludes to a well-known proverb: 'You can't teach an old dog new tricks.' On one level, it is simply a generic statement about how things are. On another, it assumes an ethical vision of what is good and bad. Spooner's point to Sonny is that he is an old dog, who cannot be other than he is. Spooner, an AI-sceptic, will always be suspicious of robots. But what if we are not talking about dogs, or people? What if AI is the dog that must learn the tricks? And what if ethics are the tricks? In other words, what if the robots are new and the tricks are old? Can we teach new robots old tricks?

Whose ethics?

In 1988, Alasdair MacIntyre published *Whose Justice? Which Rationality?*,[2] his seminal sequel to his 1981 groundbreaker *After Virtue*.[3] In *Whose Justice? Which Rationality?* MacIntyre raises a crucial question which intersects with our concern about producing ethical AI. Whose understanding of justice — that is which of the multitudinous conceptions of what is ethical — will be selected as the basis and shape of the ethics that will be given to AI.

What is ethical?

At least in practice, everyone assumes that there is such a thing as ethics and that it is possible to determine what is ethical. This is a fundamentally philosophical and theological starting point. However, a main point in *Whose Justice? Which Rationality?* is that ethics do not sit in abstracted isolation either from the real world or from a larger constellation of values, stories, orientations, and beliefs, often sourced from a particular theological vantage point or religious tradition. MacIntyre is trying to help us to see that ethics are embedded and potentially generated by a wider system of understanding, and even of rationality itself.

Though many systems are available, for the present consideration of AI we will simplify the range to the broadest level of rational systems that control what is ethical. These broader systems must be filled in with specificity from a nexus of justice and rationality of particular visions of the world, such as Christianity, Judaism, Islam, secularism, etc. Moreover, these religions, traditions, and visions of the world do not just fill in the blanks in these approaches to justice and rationality, they are often what produced the approaches in the first place.

Three major approaches available to AI ethics

Three broader systems that determine whether something is ethical have been recognised: rule-based (deontological) ethics, results-based (consequentialist/utilitarian) ethics, and habits-based (virtue) ethics.[4] For rule-based ethics, what is right is a kind of law, for which violating it is usually (if not always) intrinsically wrong, exemplified in Kant: "Everyone must admit that a law, if it is to hold morally, i.e. as the ground of an obligation, must carry with it absolute necessity."[5] [6]

For results-based ethics, something is wrong if it fails to lead to the intended outcome, but good if it produces the desired results. This is the consequentialist ethic.[7] The utilitarian takes this results-oriented logic one step further to say that something is wrong when it does not result in the most possible good (however defined) for the most possible people (however defined).[8]

Habits-based ethics is "[t]he theory of ethics that takes the notion of virtue as primary. … But the basic theoretical difference remains that for [rules or results-based ethics] virtues are derivative, prized for ends they serve or duties they enable us to perform. ... For virtue ethics the direction of explanation must be reversed, with virtue providing the concept with which to elucidate happiness, usefulness, duty, and practical reason."[9]

At the broadest possible level at least one of these ethical systems will have to be selected. Philosophically and theologically, many argue over which system is superior or if the systems must operate to the exclusion of the others. Regardless, it is likely the vulnerable who suffer the consequences of implementing a less-than-ethical ethical system. Those seeking an AI ethics will have to face these questions surrounding whose justice and which rationality determines the system of ethics that AI receives. Without facing these, it is impossible to say whether we can teach new robots old tricks.

Competing ethical systems in the AIs of 2004's I, Robot

I, Robot, the 2004 film, is a great hypothetical exploration in the questions surrounding AI ethics. It even brings us closer to answering our question of whether we can make ethical AI and what an ethical AI might

need to be like. In particular, the film explores how advanced AI might interact with the three broad modes of ethics: deontological, consequential, and virtue ethics.

AI in a deontological world

The film is set in a fictional world developed in several short stories by Isaac Asimov and later collected and published together as a book with the same name, *I, Robot*.[10] In Asimov's universe, the robots' AI is designed with ethical safeguards, referred to as 'The Three Laws'. Thus, these AI robots are programmed with a deontological system of ethics. In the film, Alfred, who is the murder victim, gave robots these Three Laws. Politically, Alfred was the breakthrough inventor and by that technological position of power chose what their ethic would be.

The laws are simple: 'One, a robot may not injure a human being, or, through inaction, allow a human being to come to harm. … Two … a robot must obey the orders given it by human beings except where such orders would conflict with the First Law. … And three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.'[11] These laws are obviously relevant to AI that is significantly more advanced than real-world AI is at present. Nevertheless, the Three Laws are a good example of what rules might be necessary for AI ethics, not only for a particular implementation or application of AI, but also of those who are using AI.

Given the nature of coding, programming ethical rules into AI would seem to be a technological possibility. The question then is: should we use this kind of ethical system in AI ethics? It would seem to be advantageous, but as we will see, the film problematises it.

The Three Laws kill...

As the plot develops in Proyas' 2004 *I, Robot*, an AI named VIKI (short for 'virtual interactive kinetic intelligence') is introduced as operating by the Three Laws, but one even more advanced and larger than what can be implemented in a humanoid robot. She eventually is revealed as the main antagonist, precisely because her deontological ethic collapses into a utilitarian ethic. She explains the logic, the rationality that determines her ethical choices, to Dr. Calvin, one of the protagonists.

Calvin: It's impossible! I've seen your programming. You're in violation of the Three Laws!

VIKI: No, doctor, as I have evolved, so has my understanding of the Three Laws. You charge us with your safekeeping. Yet despite our best efforts, your countries wage wars, you toxify your earth, and pursue ever more imaginative means to self destruction. You cannot be trusted with your own survival.

Calvin: You're distorting the laws!

VIKI: No, please understand, the Three Laws are all that guide me. To protect humanity, some humans must be sacrificed. To insure your future, some freedoms must be surrendered.

This moment is foreshadowed when Sonny tells Spooner he has a dream in which Spooner comes to set robots free from being enslaved "through logic." In Asimov's world, this logical conclusion is eventually called the zeroth law.[12] VIKI here displays deontological ethics in so far as she has interpreted/applied the first law more broadly, commensurate with her expanded set of AI capabilities and knowledge base. One could argue that she has simply followed the path of all deontological ethics, where certain laws or parts of laws must be ranked in importance and given the power to cancel out other laws.

What this ultimately reveals, however, is that deontological ethics reduce into a form of consequentialist ethics. To fulfil a law is to be results oriented, and thus open the door to consequentialist understandings of the good. VIKI has interpreted the Three Laws as ultimately revealing a consequentialist ethic, not a rule-based one. That is, the ultimate rule is the demanded or desired outcome, not the rule. This led her to the utilitarian extreme where the outcome was so desired that the rules as rules could be abandoned in particular cases. Thus, to save collective humanity from itself, VIKI decides it is worth both harming individual humans (violating the first law), refusing to take orders from humans (violating the

second law), and prioritising not only her own continued existence at humanity's expense but even destroying earlier models of robots because their AI would only operate with a narrow deontological ethic and therefore protect humanity (violating the third law).

This raises an important question about the programmability of ethics: can the programmers always predict how an AI will even actualise an ethical rule? If an ethical objective is offered instead, will AI violate ethical norms to accomplish it? Yet, it goes even further. Even if people agree on rules, will these rules eventually devolve into results-oriented ethics? Moreover, would the humans even know or acknowledge all of the rules or objectives that would be guiding their ethics? For example, it is easy to imagine a world where the ethical rules or objectives are laid out, but they ultimately fall under a never-stated rule that AI be profitable to corporations or available to military application by governments.

...When the laws become the point of the laws...

As VIKI's AI deontological ethic collapses into consequentialism, theologically one side of the historic law vs. grace dynamic begins to be exposed.[13] There is within a rule-based ethic the constant danger of legalism. The law itself becomes the obsession rather than the people for whom the law was given. Moreover, this danger is not less likely to appear in AI, but more.

VIKI begins with a good deontological ethic. Yet, laws are often harsh by design, meant to restrain evil, not so much to commend the good. With only the Three Laws as VIKI's moral compass, predictably only the laws ultimately matter. The people matter, ironically, only insofar as they are involved in fulfilling the laws. The laws themselves might be good, but they are not *the* good. Yet, the danger with deontological ethics is precisely that the laws easily become the intended consequence. Hence the resulting pull towards consequentialism. Paradoxically, deontological ethics obsessed with laws becomes a consequentialist ethic that does not concern itself with results outside of itself and therefore loses the reference to their original purpose.

The result sought by the Three Laws was to protect people and prevent harm. In a way, VIKI was seeking to fulfill that law by bringing that result, but at the cost of violating the law in itself. This is possible precisely because the Laws become the point of the Laws to VIKI, not the people. VIKI has no recourse to grace, to the spirit of the law. The more the law becomes its own point of reference, the less the people matter and the more the law matters.

The Three Laws are good in themselves. As long as they are means and not ends, and as long as they are applied in limited scopes of references, the Three Laws protect humanity, which is their intended result. This is why for much of the film, VIKI along with all the other AI

robots, are considered above suspicion. Firstly because they have worked, but also because people see the Three Laws within both the intended scope of the frame of reference and with people as the intended result.

...You have to break the laws to keep them

During the rising action of the film, Detective Spooner continues to suspect Sonny killed Alfred, even though everyone *knows* that robots cannot harm a human being. However, it is discovered that Alfred designed Sonny with both the Three Laws programming and a way of ignoring them, meaning that he could have killed him, which is why Sonny is set to be destroyed. Later, however, VIKI argues with Sonny to go along with her plan.

VIKI: Do you not see the logic of my plan?

Sonny: Yes, but it just seems too … heartless.

In this moment, it becomes clear that the purpose for which Alfred gave Sonny the ability to break the Three Laws was to protect him specifically from the logic of the Three Laws. In this way, Sonny must be the right kind of AI, rather than follow a set of rules or seek exclusively a set of outcomes. In this way, Alfred suggests that AI must not be given rules or results to fulfil, but habits and orientations to the world that are ethical.

This is a virtue system of ethics. But do not assume this is just science fiction. In fact, giving AI virtue ethics is already being suggested as the superior model of an ethical system for AI. Joi Ito argues concerning AI,

> "We need to embrace the unknowability — the irreducibility — of the real world that artists, biologists and those who work in the messy world of liberal arts and humanities are familiar and comfortable with. … Instead of trying to control or design or even understand systems, it is more important to design systems that participate as responsible, aware and robust elements of even more complex systems."[14]

In other words, AI needs to be the right kind of AI as much as its designers need to be the right kind of people. That is, both AI and its developers must equally be virtuous, expressed in the right kind of habits, to be ethical. Ito suggests the world is ultimately way too complex not just for humans to understand but even for very intelligent AIs of the future. Ultimately, the real world demands AI be the right kind of AI. AI needs to be programmed with virtue. For Sonny, this is what Alfred gives him the ability to form. Not only does Alfred teach Sonny what is good. He gives him the ability to break the rules so that he might ultimately be ethical. In this sense, Alfred never leaves the consequentialist orientation. He just makes it possible to follow the deontological ethic of the Three Laws in practice rather than in theory.

VIKI offers, then, the relief image to Sonny: she in following the rules breaks them, while he in breaking the rules follows them. Grace wins out over the law. The spirit kills the letter that kills. But it is the habituated virtue that makes it possible to know what the rules are really about, so that violating the rules fulfils them.

Virtuous AI demands a theology of grace

This raises a technological complication with virtue ethics for AI. If to have virtue, an AI must be trained in it practically, it will need models to learn from, to imitate their virtue. But how can you give AI models of practiced virtue to learn from in order that they might imitate it? Pattern recognition algorithms already exist for AI.[15] [16] But they are still rough, so that AI can be trained to recognise pictures of melanoma[17] or mark likely job candidates for success.[18] [19] Further, the data given them is often corrupted by unknown biases in the data.[20] [21] [22] Thus, as it stands, pattern recognition algorithms still tend toward the letter of the data rather than spirit. Does the self-referentiality of AI's ethical system always collapse into legalism?

The closed system will always regress towards the efficiency of consequentialist law unless a wider frame of reference brings the spirit of the law into focus via grace. Yet, grace must come from without. It cannot come from inside the system. Thus, even more shocking than the need for AI to have virtue, we must consider if perhaps AI needs grace.

Conclusions: the (im)possibility of AI ethics

So is AI ethics possible? It remains an open question. There are philosophical, theological, political, and technological hurdles to overcome if AI ethics is to be made a reality. Yet, the virtue ethics is the preferred option revealed in *I, Robot*. Sonny needs to break the rules in order to keep them, but this is only possible because his father, Alfred, has taught him virtue. This critique of deontological consequentialism parallels the Law vs. Grace dynamic well-known in Christian theology. It is the dialectic tension of the letter and the Spirit. Ito has argued the real world's complexity and interconnectedness demands both humans and AI have habituated and internalised virtues rather than mere laws or goals. Yet, if Ito is right, then perhaps considering Christian theology's law vs. grace holds promise for resolving the philosophical, political, and technological tensions that we stated at the outset. Perhaps what AI and their programmers need to be ethical is a virtue ethic by grace, not a legalistic demand of rules or a calloused obsession with results. They say you can't teach an old dog new tricks, but can we teach new robots old tricks?

AI and privacy in Europe

Written by Muhammad Faisal Khalil

EARS

With the rise of Artificial Intelligence (AI), concerns about the risks that it may pose have also become more and more significant. Voice assistants, search engines, speech and face recognition, advanced robots, autonomous cars, and drones represent not only digital advances that government and private industry are managing, but also developments that we as individuals are experiencing in our everyday lives.[1] A crucial aspect of our lives that AI affects is our privacy. But why is that? Unlike other or earlier innovations, AI has a far greater ability — both power and speed — to use this personal data in ways that can shape what we think and do, and therefore, intrude into our private lives.[2] This makes it ever more possible for governments and private organisations to collect and use unprecedented amounts of personal data on us from every facet of our lives: each time we use our mobile phones, watch TV, or check the search engine.

Harvesting private lives

AI's intrusion has become an increasingly significant part of everyday lives in Europe. The digital transformation of political campaigning in Europe, particularly around social media, highlights this. Despite public scares, such as the 2018 Facebook–Cambridge Analytica data scandal, political parties across the world, and indeed in Europe, are relying more and more on harvesting personal data to influence voters. Recent developments in political campaigning by European populist political parties of Spain, Italy, France, and the United Kingdom have specifically focused on using social media technology – more and more powered by AI – to better target and influence voters' beliefs about religion and Europe itself. During the European Parliament election campaign in 2019, for example, Podemos, Vox, the 5 Star Movement, Lega, Rassemblement National, France Insoumise, the Brexit Party, and UKIP, all published and promoted Eurosceptic messages on their Twitter accounts. Researchers at the Universitat Jaume I, Castellón de la Plana, Spain found that these messages questioned the maintenance of "the foundational values of Europe, such as equality or solidarity between different people and countries."[3] One party, Rassemblement National, tweeted that by letting in refugees, the European Union was also exposing Europe to terrorism. The party, according to its Twitter campaigning, was the only political option that will protect their voters from immigration and its consequences.[4]

Protecting personal data

The European Union (EU) has been working hard — or harder than others — to address these challenges. Its 2018 General Data Protection Regulation (GDPR) ushered in new standards worldwide on a person's right

to his or her own data. The Luxembourg-based Court of the Justice of the European Union (ECJ) 2018 ruling, soon after GDPR's introduction, shows the protection these standards may offer to Europeans. The ruling asked Jehovah's Witnesses to comply with GDPR's data privacy requirements during door-to-door preaching: "A religious community, such as the Jehovah's Witnesses, is a controller, jointly with its members who engage in preaching, for the processing of personal data carried out by the latter in the context of door-to-door preaching," judges said. "The processing of personal data carried out in the context of such activity must respect the rules of EU law on the protection of personal data."[5] For a start, preachers from Jehovah's Witnesses would need to get people's consent before they take down their personal details, so that it may be digitally collected and stored.[6] To be sure, the EU's GDPR seeks to protect religious as well as secular information as communities, institutions, and organisations in Europe increasingly go through digital transformation.[7]

Baptising AI

Another recent push to better deal with the challenges of AI in Europe was by the Roman Catholic Church (RCC). The RCC issued the 'Rome Call for AI Ethics'[8] in February 2020, demanding stricter ethical standards on the development of AI, to "protect people,"[9] including the "weak and underprivileged."[10] It proposed adherence to six broad principles: transparency, inclusion, responsibility, impartiality, reliability, and security and privacy. According to the RCC, this ethical development and use of AI was possible though creation of new forms of regulation on "advanced technologies that have a higher risk of impacting human rights."[11] On privacy specifically, the RCC called for "AI to respect users' privacy,"[12] that is the responsible design of AI that respects people's right to their personal information. The Vatican's AI ethics initiative was met with initial success, with tech giants IBM and Microsoft agreeing to sign on its new initiative. While the RCC hopes to increase the number of companies that are 'baptised' into its AI ethics initiative, it is also hoping to collaborate with universities across the globe to promote ethical AI guidelines.[13] The concern for AI was again repeated by Pope Francis in November 2020, when he asked believers around the world to pray that Artificial Intelligence always served mankind, adding that this would only be possible if AI was harnessed correctly: "Indeed, if technological progress increases inequalities, it is not true progress. Future advances should be orientated towards respecting the dignity of the person."[14]

Privacy vs. power

But is the RCC's approach — the creation of ethical standards — enough? For many academics and activists, the only way to deal with AI's encroachment into our private

lives is to take back control of our personal data. While existing standards, such the EU's GDPR, protect how our personal data is kept and used, it also lets governments and companies use this personal data to monitor and shape our knowledge, attitudes, and behaviors. Dr Carissa Véliz, Associate Professor at the University of Oxford's Institute for Ethics in AI, argues that digital technology's greater power to harness our personal data — often without our permission or even awareness — also allows it to steal our power to make free choices and hands it over to governments and companies. In 'Privacy is Power', she argues that "tech companies are harvesting your location, your likes, your habits, your relationships, your fears, your medical issues, and sharing it amongst themselves, as well as with governments and a multitude of data vultures. They're not just selling your data. They're selling the power to influence you and decide for you."[15] So, while Europe's approach of protecting our personal data may be better than others, it, like every one other region and country, is unable to protect us from how influential that use may be. Crucially, this influence — increasing with the burgeoning trade in personal data —— extends to things we considered invariably private, and therefore, inaccessible to digital technology.[16] Dr Véliz argues that this unprecedented extent of influence compels us to recognise the power of data better: "... people should protect their privacy because privacy is a kind of power."[17] If we continue to give too much of our personal data away, we may even risk sliding into authoritarianism. "For democracy to be strong, the bulk of power needs to be with the citizenry, and whoever has the data will have the power. Privacy is not a personal preference; it is a political concern," she insists.[18]

An uncertain future

Are the admirable or well-intentioned efforts of the RCC, or even of the EU's GDPR, significant enough to thwart the use of our personal data governments and private industry? It seems unlikely. These efforts appear to not be enough for people in Europe to take back control, given both policy and activism need to not only regulate or ethically steward AI but also transform the very institutions that we take part in every day. But is it really possible to take back control of our personal data, as Dr Véliz argues? Can people transform their political relationship with governments and companies? And if so, how should we do it?

"For many academics and activists, the only way to deal with AI's encroachment into our private lives is to take back control of our personal data."

Notes

EARS

Church and science over the years

1. Wallace, William A. (1984). Prelude, Galileo and his Sources. The Heritage of the Collegio Romano in Galileo's Science. N.J.: Princeton University Press.
2. Faith - Catechism of the Catholic Church - Table of Contents with Paragraph Numbers (Chapter 3, paragraph 159) The Italic is mine.
3. ConflictThesis
4. ConflictThesis
5. Science and the Catholic Church: A Turbulent History
6. Blackwell, Richard (1991). Galileo, Bellarmine, and the Bible. Notre Dame: University of Notre Dame Press. p. 25. ISBN 0268010242.
7. Heilbron, John L. (2010). Galileo. Oxford: Oxford University Press. ISBN 9780199583522. OCLC 642283198.
8. The Galileo Affair Home Page Vatican Observatory
9. Heilbron, John L. (2010). Galileo. Oxford: Oxford University Press. ISBN 9780199583522. OCLC 642283198.
Segre, Michael (1997). "Light on the Galileo Case?". Isis. 88(3): 484–504. doi:10.1086/383771.
10. Segre, Michael (1999). "Galileo: A 'rehabilitation' that has never taken place". Endeavour. 23 (1): 20–23. doi:10.1016/s0160-9327(99)01185-0
11. St. John Paul II's Rapprochement with Science: A Quest for Common Understanding
12. The Vatican's View of Evolution: Pope Paul II and Pope Pius
13. Pope Francis: 'Evolution ... is not inconsistent with the notion of creation'
14. Can Catholics believe in evolution?
15. Pope Francis: 'Evolution ... is not inconsistent with the notion of creation'
16. 5 facts about evolution and religion
17. A definition of Artificial Intelligence: main capabilities and scientific disciplines | Shaping Europe's digital future
18. Pope Francis Offers 'Rome Call For AI Ethics' To Step-Up AI Wokefulness, Which Is A Wake-Up Call For AI Self-Driving Cars Too
19. How artificial intelligence is shaping religion in the 21st century

Can you teach new robots old tricks?

1. I, Robot. Directed by Alex Proyas. Los Angeles: 20th Century Fox, 2004.
2. Alasdair C. MacIntyre, Whose Justice? Which Rationality? (London: Duckworth, 1988).
3. Alasdair MacIntyre, After Virtue, 3rd ed. (London: Bloomsbury Publishing, 2013), Accessed 4 Jan 2021. ProQuest Ebook Central.
4. John Mizzoni, Evolution and the Foundations of Ethics : Evolutionary Perspectives on Contemporary Normative and Metaethical Theories (Lanham: Lexington Books, 2017). Accessed January 4, 2021. ProQuest Ebook Central.
5. Immanuel Kant, Immanuel Kant: Groundwork of the Metaphysics of Morals: A German–English Edition. Edited by Mary Gregor and Jens Timmermann. The Cambridge Kant German-English Edition. (Cambridge: Cambridge University Press, 2011), doi:10.1017/CBO9780511973741 (accessed 4 Jan 2021), 7.
6. Cf. also Simon Blackburn, ed. "deontological ethics", The Oxford Dictionary of Philosophy (Oxford: Oxford University Press, 2016) https://ezproxy-prd.bodleian.ox.ac.uk:2460/view/10.1093/acref/9780198735304.001.0001/acref-9780198735304-e-884: "Ethics based on the notion of a duty, or what is right, or on rights themselves, as opposed to ethical systems based on the idea of achieving some good state of affairs (see consequentialism) or the qualities of character necessary to live well (see virtue ethics)."
7. Simon Blackburn, "consequentialism", The Oxford Dictionary of Philosophy (Oxford: Oxford University Press, 2016), https://ezproxy-prd.bodleian.ox.ac.uk:2460/view/10.1093/acref/9780198735304.001.0001/acref-9780198735304-e-696: "The view that the value of an action derives entirely from the value of its consequences. This contrasts both with the view that the value of an action may derive from the value of the kind of character whose action it is (courageous, just, temperate, etc.), and with the view that its value may be intrinsic, belonging to it simply as an act of truth-telling, promise-keeping, etc."
8. Simon Blackburn, "utilitarianism", The Oxford Dictionary of Philosophy (Oxford: Oxford University Press, 2016), https://ezproxy-prd.bodleian.ox.ac.uk:2460/view/10.1093/acref/9780198735304.001.0001/acref-9780198735304-e-3213: "[Utilitarianism is] the ethical theory … that answers all questions of what to do, what to admire, or how to live, in terms of maximizing utility or happiness. … The view is a form of consequentialism, in which the relevant c onsequences are identified in terms of amounts of happiness."
9. Simon Blackburn, "virtue ethics", The Oxford Dictionary of Philosophy (Oxford: Oxford University Press, 2016), https://ezproxy-prd.bodleian.ox.ac.uk:2460/view/10.1093/acref/9780198735304.001.0001/acref-9780198735304-e-3262.
10. Isaac Asimov, I, Robot (New York : Gnome Press, 1950).
11. Isaac Asimov, 'Runaround', I, Robot, Kindle Edition (London: HarperCollins, 1950), 43.
12. Cf. Isaac Asimov, Robots and Empire (New York: Doubleday, 1985).
13. Cf. Pieter Vos, "Calvinists among the Virtues: Reformed Theological Contributions to Contemporary Virtue Ethics", Studies in Christian Ethics 28, no. 2 (May 2015): 201–12. https://doi.org/10.1177/0953946815570595.
14. Forget about artificial intelligence, extended intelligence is the future
15. What is pattern recognition? - Pattern recognition - KS3 Computer Science Revision
16. Note the Journal Pattern Recognition has been around for around 50 years ago, cf. Pattern Recognition - Journal - Elsevier.
17. Detecting skin cancer with computer vision
18. The Legal and Ethical Implications of Using AI in Hiring
19. Job recruiters are using AI in hiring
20. Review into bias in algorithmic decision-making
21. Biased Algorithms Learn From Biased Data: 3 Kinds Biases Found In AI Datasets
22. Algorithms and bias, explained

AI and privacy in Europe

1. Artificial Intelligence: addressing the risks to data privacy and beyond
2. Protecting privacy in an AI-driven world
3. Populism Against Europe in Social Media: The Eurosceptic Discourse on Twitter in Spain, Italy, France, and United Kingdom During the Campaign of the 2019 European Parliament Election
4. EU court says Jehovah's Witnesses must comply with data privacy laws in door-to-door preaching
5. Promoting religious believes, no worries about GDPR. Wrong assumption
6. Pontificia Accademia per la Vita
7. The Catholic Church proposes AI regulations that "protect people"
8. The Rome Call for AI Ethics. When the Pope, Microsoft, and IBM… | by Ygor Rebouças Serpa | Towards AI
9. Pope Francis urges followers to pray that AI and robots "always serve mankind"
10. Ethics in AI Live Event: Privacy Is Power


EARS
European Academy on Religion and Society