

Humanistic Toolkit

GIFT (Artificial Intelligence, Philosophy and Technology Research Group)

Dr. Karina Pedace

Dr. Tomás Balmaceda

Dr. Diana Pérez

Dr. Diego Lawler

Prof. Maximiliano Zeller Echenique

GuIA.ai

In collaboration with

fAIR LAC



PART 1

- Notes for philosophical reflection on artificial intelligence
- Some philosophical reflections on the design of artifacts and technological systems

In this work we present a general framework for considering artificial intelligence and its applications from a humanistic perspective. The purpose of this work is to analyze these developments by drawing attention to the social and human environment in which these new information technologies arise and are applied, as well as the ethical implications raised by these new developments. The work is divided into two parts.

In the first part, we present a series of philosophical reflections aimed at clarifying the nature of artificial intelligence, its relationship with human intelligence as well as the various types of developments that have occurred in this area with the issues and challenges they pose. In order to understand more fully the ethical implications that the development of these technologies implies, we also present a general philosophical framework regarding the design of artifacts and technological systems in which the various human actions and practices employed are revealed, as well as the different levels of nested ethical implications in these developments.

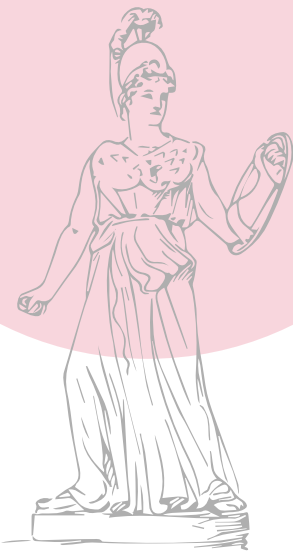
In the second part, we focus on some of the most discussed ethical problems in recent literature with the greatest impact on the development of AI technologies: the bias problem, the data privacy issue, system transparency, responsibility allocation in the design and use of these systems, security, and finally, their relationship with well-being, with human rights and with human society as a whole.

PART 2

- Biases
- Privacy
- Transparency
- Responsibility
- Security
- Well-being, human rights, politics and technology

PART 1

■ Notes for philosophical reflection on artificial intelligence.



Since Ancient Greece, the idea that human beings can be described as animals with logos, that is, with reason and speech has been maintained. We consider ourselves "rational animals" and we understand that what distinguishes us as a biological species from other animals and machines (automata, robots) is our intelligence, our rationality and our speech. There seem to be problems that only humans can solve (mathematical theorems, games such as chess or Go) and behaviors that only human beings can perform (write a poem, start a dialogue with another human being, suffer for love). Both our theoretical intelligence (our ability to understand the world and accumulate knowledge) and our actions (our ability to solve the practical challenges that the world and human society pose before us) are understood to be the result of our ability to solve problems rationally.

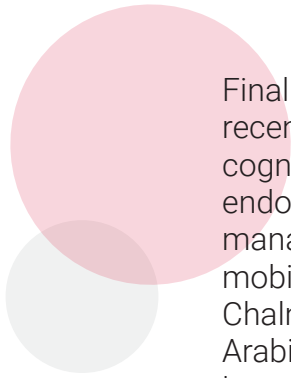
In the middle of the 20th century our human self-image was questioned by the British mathematician and cryptographer Alan Turing (1950) in the provocative article "Can machines think?". In this work, Turing, echoing the developments in logic and computer science of the time, to which he himself remarkably contributed, challenged the Cartesian idea that banished machines from the possibility of engaging in a dialogue as if they were human. The "imitation game" - as conceived by Turing - would allow a machine (an automaton programmed to engage in dialogue with a human) to "trick" a human by mistaking the machine for another human. Thus, artificial intelligence arises under the premise of building an artifact capable of executing intelligent behaviors indistinguishable from human behavioral patterns.

This project can be considered in a less ambitious or narrow manner (narrow AI) if it deals with the search for systems that can perform a particular task intelligently. Alternatively, it can be viewed in a more ambitious or general manner (general AI), if the intention is to create an artifact that in all respects acts (and thinks) in a humane manner (like the replicants from the Blade Runner movie). Clearly, this second project is still relegated to the field of science fiction; hence, thereafter in this work we will focus on the narrowed down project, namely artificially replicating certain specific human capacities.

If we stop to reflect on what is involved in the notion of "intelligent behavior", we will quickly notice that it contains a normative idea. Indeed, what counts as intelligent or rational behavior depends not only on the behavior actually performed by people, but on whether those behaviors are supported by principles indicating that it is the best behavior or the most appropriate for the occasion (Aristotle, 1973; 2010; Raz, 1979; Pedace, 2017). The behavior displayed by an individual is intelligent in light of certain reasons that led them to act as they did: eating a four-cheese ravioli dish can be rational if we do it because we haven't eaten in a long time, because we are hungry, because we want to stay alive, and because we like pasta and cheese, but it can also be irrational, or not very correct, if we have already eaten barbecue, if we are overweight, if we have high blood pressure, high glucose, etc. Concurrently, a person's beliefs can be rational if they are based on verifiable evidence, if they follow logically from other well-founded previous beliefs, or irrational if we simply echo a neighbor's gossip or some fake news that appeared on our social networks.

Reflections on what is considered intelligent behavior and what is not intelligent entails establishing rationality parameters that quickly lead us to contemplate a novel possibility from which the idea of artificial intelligence is nourished. It seems that not only could an artifact capable of imitating effective intelligent human behavior be built, but a more perfect machine than any flesh-and-blood human could also be built, a machine that could solve human problems better than humans themselves without making mistakes, that is, without ever deviating from the norms of rationality. Creating machines that perform some human task in a way that surpasses the best known human has been a driving force of artificial intelligence, a project that crowned Deep Blue as world chess champion by defeating Kasparov or AlphaGo by defeating Lee Sedol, the world champion of Go.





Finally, an additional and even more provocative idea can be found in recent literature. There are good reasons to believe that human cognition / intelligence depends not only on our biological endowment but also on the cultural niches that human beings have managed to build by creating artifacts such as the alphabet and mobile phones to enhance our cognitive abilities (Clark and Chalmers 1998; Clark 1997; 2008). The creation of the alphabet, the Arabic notation for natural numbers and the creation of artificial languages in general for logic, arithmetic and computer science are tools that, coupled with our biological endowment, have enabled human beings to perform feats that were unimaginable 40,000 years ago (or even 150 years ago!). We can then ask ourselves if it would not be possible to generate a new cultural niche with AI development that allows human beings to change their cognitive abilities in a drastic, and as of yet unimagined, manner. Are we not now in a situation that is similar to the one we were in before the existence of the alphabet? The challenge of combining these AI developments with existing human capabilities, by enhancing human capabilities (human enhancement), allows us to think that perhaps we are facing the possibility of generating "superhuman" or "transhuman" beings, that is, beings that break down the barriers of what our biological endowment together with the currently existing cultural molding allow us to do.

Discussion exercise



Is there something distinctively human?

More than a few people feel intimidated and disturbed by the possibility of increasing our human abilities or creating new ones thanks to technology. The root of this discomfort seems to be the idea that there is such a thing as "human nature" that cannot be altered or polluted ... What do you think? What is it that makes us men and women, setting us apart from other animals and objects in the world? As we mentioned at the beginning of this section, it was believed for centuries that our reasoning ability and language made us unique. Now that machines can challenge us for this achievement and that some definitions of intelligence open the way for us to consider various animals as intelligent... Is there something distinctively human that no other being shares?

We can see in this brief presentation that there are several different projects coexisting in the AI field, and not all of them coalesce in the same place. Indeed, there are at least three different purposes surrounding the construction of AI systems.

AI Projects



- 1** To generate systems that act and think as human beings actually think nowadays (with defects and virtues).
- 2** To generate systems to perform activities as an ideal human would, that is, as a perfect human would do who never made mistakes, who was rational to the utmost, etc.
- 3** To generate systems that enhance existing capabilities by creating systems composed of humans + machines, capable of doing things beyond the ideal perfection we conceive today.

All of these projects involve a series of hidden normative principles. On the one hand, the generation project of artifacts that act and solve problems just as effectively as existing human beings do, which inherit human tics, biases and defects. A system that wanted to act like a human should be wrong from time to time, it should take longer to answer a complex question than a simple one, it would show biases in judging human beings, prejudices, etc. Otherwise, it would be very easy to realize that a machine is answering us instead of a human being.

It seems, however, that the interesting promise shown by AI is the generation of systems that allow us to improve human societies, that is, that act as a perfect human being and not exactly like us. If we want to improve justice, for example, we should create an AI system that is a perfect judge, that only fails to administer justice while doing it the right way; or a perfect doctor who is never wrong in their diagnoses if we want to improve current medical practice. Obviously, these ideal systems inherit the normative canons of their idealism. If a system is logically rational, it should respond by appealing to logically valid reasoning, it should act in

a way that maximizes the utility of its actions, etc. If we want to build a perfect judge, we should be clear on what exactly is the fairest decision in each case ... but this seems to be quite removed from our human possibilities. Ultimately, the best way to administer justice is an issue that human beings will probably never agree on (nor, unfortunately, are there perfect diagnoses: there is still a big lack of knowledge of the human body and its psychosocial environment).

Finally, if we want to "improve" human beings in such a way that they exceed the cognitive possibilities of existing humans, we must be clear about which things are better than others and to what extent. Becoming superhuman means valuing one thing better or worse than another or one situation before another, and it is not always evident what a better human life consists of: living longer without certain pleasures or fewer years with more pleasures; fight for our homeland on the battlefield or desert and hide from war, etc. (And probably, once again, not all humans have the same answer to this question).

Discussion exercise

What is a "good life"?



The question of "What is a good life?" is almost as old as philosophy: it was already mentioned by the first authors we know of and remains open to this day, since there does not seem to be a single answer or opinion that satisfies everybody. Despite its longevity, it is still a hot topic: what it is that we will consider a good life will be the main focus when it's time to rethink our ethical decisions regarding artificial intelligence.

As we can see, the development of AI systems raises a host of ethical questions and, more generally, philosophical questions both pertaining to the practical consequences these developments could have on human lives as theoretical questions about the extent to which they alter the self-image that we humans have of ourselves. Our privileged position in the world has been called into question.

Let's see where we stand today (2019) as regards AI developments and list **some of the achievements that make up our current human niche:**



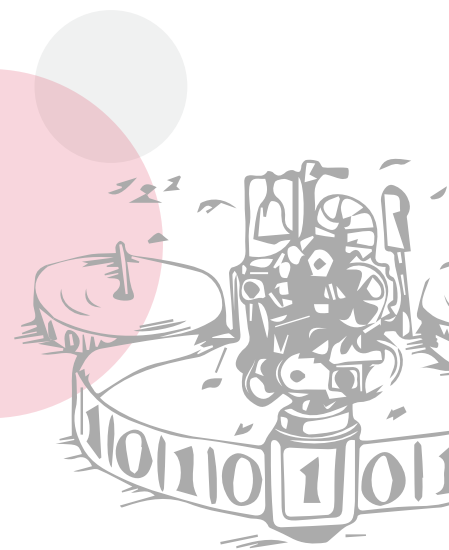
- We can dialogue (chat) with machines - and we usually do it on Internet sites where chatbots offer us services, including medical "services", legal consulting services, etc.
- We can count on (fairly reasonable) machine translations of portions from one natural language to another natural language.
- Some machines can recognize faces and fingerprints (at country borders, in security cameras on the street and on public transport, in certain buildings, when unlocking our phone's screen...).
- Some machines can recognize and label objects in our environment.
- There are autonomous vehicles (i.e., that allow a human to be moved from one place to another without a human "driving" the vehicle).
- There are also weapons that operate autonomously (i.e., they "decide" what objective to hit with their missiles without human intervention).
- There are consulting firms that offer identification services for individuals who are accessible targets for political or commercial manipulation.
- Almost all social networks and streaming platforms offer us advertising and news; they provide us with customized job opportunities and services according to our interests and preferences.
- There are complex machines that play games better than the human world champion in each specialty: Deep Blue defeated Garry Kasparov and AlphaGo defeated Lee Sedol.
- We can move about in cities only with the help of our mobile, which is connected to the Internet.
- There are systems that prove propositional logic theorems (and we know that it is not possible to do the same with arithmetic).
- There are countless automatic machines that replace humans in factories and service companies.
- There are traffic lights regulating the flow of traffic efficiently thanks to AI.
- We have apps that promise to help us find our ideal partner.

- We have automatic spell-checkers for our messages, automatic replies for our emails, etc.
- There are machines like Siri and Alexa that recognize our voices, understand our requests and "obey" us by turning on lights, looking for data in our agenda, writing emails for us, etc.

Not all of these AI developments share the same algorithms, that is, not all of them respond to a single mathematical model. We are not going to dwell on details, but we need to mention some general issues to clarify some of the tools we propose using to consider AI in a humanistic manner.

The first AI "programs" were based on the idea of the "Turing machine". As we already know, it is a very simple algorithm (an ideal/logical machine, not a material effective machine) that includes a series of commands, which are coded in a table where all the system functions or the transition rules are specified. Thus, a Turing machine is capable of executing a multiplicity of tasks to the extent that these transitions from one state of the machine to another can be serially applied. In these models we can see that the rules followed by the system reproduce normative canons about what should be done before each given information or situation by applying the rules serially and recursively in order to arrive at the answer to the problem posed. These serial machines are very powerful to perform certain tasks but have serious limitations for others (since its inception, the famous problems regarding the framework and computational explosion pointed out restrictions to this model).

But AI development has quickly moved away from these serial machines. Most of the AI developments involved in the aforementioned examples are based on other types of algorithms that allow patterns to be extracted from large amounts of data (usually called "machine learning" or "automated learning"). These mechanisms start from a more or less large and more or less structured series of data ("past experiences") from which pre-existing patterns are discovered, allowing them to predict what will happen when new data appears or perform several tasks depending on the accumulated past "experience" of the "learning" performed. The learning process varies; it can be guided or automatic, that is, there may be humans who correct the progress in learning – teachers - or the machine can learn by interacting with itself; it could be active or passive, that is, it can occur in the midst of a constant interaction with the environment that provides feedback for the learning process, or it can be purely observational, i.e., the machine acquires information



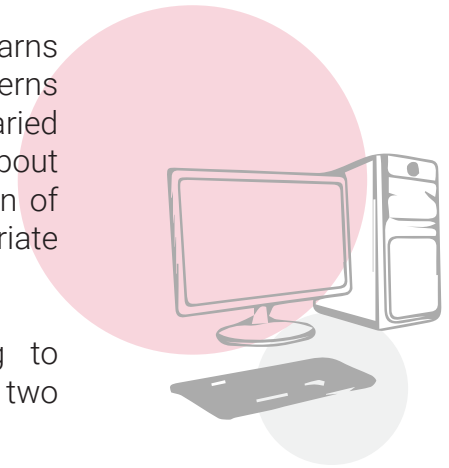
before acting; it can be oriented to an end, where the machine learns to perform a specific task or simply to find unexpected patterns among pre-existing data (data mining). The algorithms can be varied (Shalev-Shwartz and Ben-David, 2014) but, in all cases, it is about "educating" a program to generate a model of a certain domain of reality, which will allow the system to carry out the appropriate actions.

Without going into excessive technicalities, and according to Mittelstadt et al. (2016), it should be noted that there are at least two meanings for "algorithm":

- From its formal characterization, it is considered a mathematical construct for the purposes of achieving a given purpose under certain provisions.
- From its specific use, it refers to the things that must be implemented and executed to carry out an action and obtain certain effects.

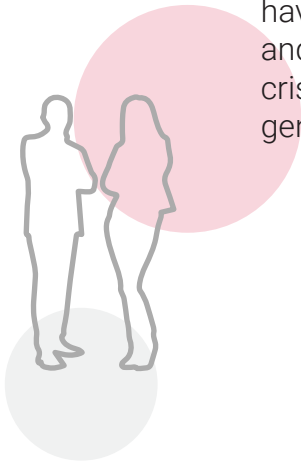
A fully configured algorithm thus incorporates the mathematical abstract structure that has been implemented in a system to analyze tasks in a particular domain. From now on, when we talk about "algorithms" we will be referring to this second sense, that is, to an algorithm made on a specific artifact that is part of our material world. Ethical reflection on algorithms seems to crucially require that we pay attention to the way in which they are implemented and executed in computer programs, software and information systems. In particular, the current scene is dominated by a particular class of algorithms that make decisions determining, for example, the best action to take in a given situation, the best interpretation of data, etc. Such algorithms increase or replace human analysis and decision making, often due to the scope or scale of the data and the rules involved. Nevertheless, algorithms question us ethically not only by virtue of the scale of analysis and the complexity of the decision-making, but also the opacity of the work they perform and its impact on our lives is a reason for special reflection.

As we said before, automated learning systems are characterized by the ability to define or modify decision-making rules autonomously. A machine learning algorithm applied to classification tasks, for example, typically consists of an apprentice producing a classifier with the intention of developing classes that can be generalized beyond the training data (Domingos, 2012). As we pointed out, this learning can be supervised (via hand-labeled inputs) or unsupervised (the algorithm itself defines the models that best fit to make sense of the set of inputs). In both cases, the algorithm defines the decision-making rules to handle new inputs. Consequently, the human operator does not need to understand the reason behind the decision-making rules produced by the algorithm (Matthias, 2004). However, it is a desideratum of AI designers that transparency does not completely disappear in this process, since algorithms that are poorly predictable or inexplicable are difficult to control, monitor or correct (Tutt, 2016). As Mittelstadt et al. (2016) point out, transparency is often naively treated as a panacea for ethical issues that emerge from new technologies.



In this way, we see that the data from which the system is based is essential for the system to learn, and what has been learned is in direct relationship with the data that we provide from the start and with the corrections made throughout the learning process. Thus, if the system starts from a previous set of unfair court rulings, it will generate new unfair rulings. If it starts with biased data (for example, data about the preferences of men and women in our current patriarchal society), it will generate biased advertising offers, reinforcing in this way the original bias: brooms will be offered to women and cars to men. We will also be shown more frequently the opinions and photos of friends to whom we gave a "like" than those from friends who never receive our comments... And the system will reinforce our old opinions by showing us the opinions of those who agree with us (confirmation bias).

Developments like those already exemplified are highly widespread in our current societies, and both their development and use raise countless ethical issues, many of which are novel and urgent. The speed with which certain systems and applications involving AI succeed each other and transform our daily lives is remarkable. In this work, we offer a series of tools originating in philosophy (and from the humanities in general) with a view to generating ethically informed critical thinking in those people who in some way or another have to make decisions that concern AI development and/or its application: designers, developers, companies, public policy managers, even parents who have to decide what they allow and what they prevent their children from doing, and, why not, users in general, citizens who live in this new cultural niche crisscrossed by AI. But first, we have to fully understand the process of generating new technological systems and their insertion in human societies.



■ Some philosophical reflections on the design of artifacts and technological systems

There are three fundamental questions to address technological design:

The **first** is the question regarding the agency: who or what is the designer?

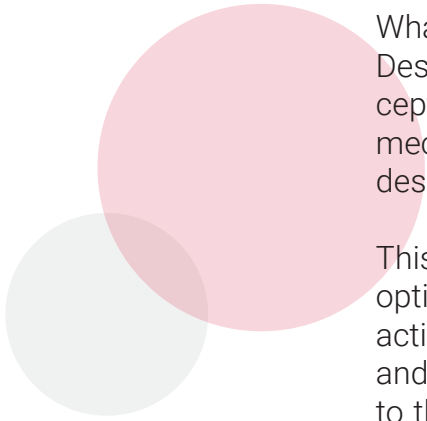
A **second** question corresponds to the design method: how is it designed?

A **third** question is basic in nature and is related to the product of the design: what is a designed object or system?

The answer to the first question is apparently simple: the design agent is a human, rational, deliberative and far-sighted designer. As regards the second question, it should be noted that the design method is based on decision-making procedures that include foresight of the practical uses of the designed entity and its consequences. Finally, the result of the design process is an artificial artifact or system that fulfills a mediation function between human beings and their environment. Once it has been decided which design will be made, the artifact or technological system is produced and these become part of our environment, proceeding to shape our daily life.

However, in the human practice of designing and producing an artifact different dimensions involving relevant ethical issues coexist. Let's consider each one of them.





What does designing an artifact or a technological system mean? Design is a human activity that consists in performing a cognitive-conceptual operation by means of which models of material structures and mechanisms are created that perform the functions proposed by designers in order to solve a situation that is posed as a problem.

This cognitive-conceptual operation includes representations of options of material structures and mechanisms, as well as courses of action, deliberation on those courses of action, on material structures and on mechanisms, establishing considerations and choosing some to the detriment of others according to a set of previously formulated values and objectives.

In the design process, faced with the representation of any technological system or artifact, there is a possible alternative design. Therefore, design is basically a deliberative area that is closed with a decision on which design will be adopted and carried out effectively and materially. Thus, when one design is chosen voluntarily over another, designers are responsible for that choice. There is a capacity, regarding the creation of designs for technological systems and artifacts, and an obligation associated with that capacity, regarding taking responsibility for the choice of design and for the consequences of its introduction into the human environment.

Choosing a design for an artifact or a technological process to be produced and introduced into the real world is one answer to the question of what the appropriate solution is to a given problem.

Discussion exercise



What does a "suitable solution" mean?

How do you determine that a solution is a "suitable solution"? This question is important because it leads to questions about the way in which technological decisions are made, according to which dimensions intervene in the creation of an artifact or technological system, and the way in which they are valued in the decision made in relation to the problems they solve and with the modifications they introduce in the world.

In the deliberative context of the design, intrinsic and extrinsic assessments are involved. Intrinsic assessments are strictly technological and extrinsic assessments are concerned with the human purposes incorporated into the artifacts and technological systems.

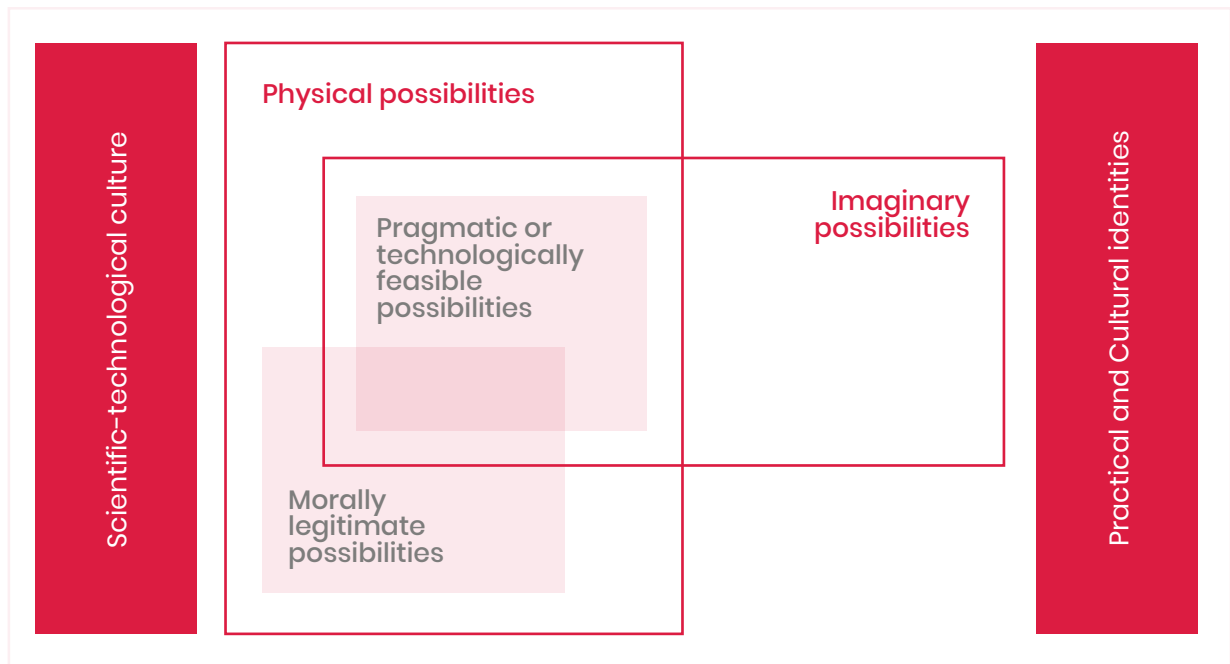


Figure 1. Design, decisions and evaluative contexts (adapted from Broncano, 2000)

- Intrinsic assessments are linked to the effectiveness and efficiency of artifacts and technological systems; to their performance, which should be achieved with the least possible waste of resources (material, economic, etc.) and of unintended consequences (killing flies with cannon shots is effective, but it has multiple unintended consequences).
- Extrinsic assessments are related to the meaning of an artifact or technological system being an adequate solution to a problem posed. This does not only depend on the artifact or system being effective and efficient; on the contrary, it depends on how it is inserted into people's environment and in a given society, and on the assessment carried out by different social agents (designers, producers, users, etc.). This is a contextualized assessment, where the values related to the way in which one wants to live count more than the strictly technological aspects. This assessment is strongly associated with the particularities of social groups as well as with their practical identities and, therefore, with their purposes, interests and points of view about the world. For example, a social group driven by a strong ecological conscience and commitment may reject the installation of a nuclear power plant in their region, despite the fact that this would be the most technologically appropriate option to solve energy supply in these socioeconomic and environmental conditions.

By delimiting these two types of assessments at stake, it becomes clear that the technical solution regarding intrinsic assessments is not all that needs to be explained. This is the old dream of positivist neutrality, which means that it is possible to sharply demarcate epistemic values such as efficiency, truth, etc., from deep ethical values such as those related to what we understand, for example, of the ideal of "human flourishing". Indeed, the merely intrinsic does not resolve extrinsic issues related to the way we want to live. Thus, it is evident that technology requires ethical reflection. In this paper we focus on the ethical issues raised, in particular, by the development of technological systems involving AI developments. We will present below the most debated ethical issues now made explicit within the humanistic framework proposed so far. Our intention is to provide the reader with a "humanistic toolkit" to rethink AI and the multiple challenges presented to us by said AI.



PART 2

■ Biases

As we have seen, algorithms are loaded with values. The operational parameters are specified by developers and configured by users who have certain desired results in mind, who privilege some values and interests over others. However, operating within accepted parameters does not guarantee that the resulting conduct is ethically acceptable. This is observed, for example, in algorithms creating profiles that discriminate against different individuals and/or social groups, even when this is not the designers' intention.

When algorithms draw conclusions from data, they are processed by means of statistical inferences or machine learning techniques that lead to the production of probable knowledge (and inevitably, uncertain knowledge; in fact, there are specific theories that deal with the characterization and quantification of this uncertainty). In this way, we find ourselves before the recognition of certain epistemic limitations. In this sense, it is necessary to explain some reflections regarding the evidence we use when making these developments and applications (cf. Mittelstadt et al., 2016).

Algorithms process data and are, therefore, subject to the restriction where the output can never exceed the input. While Shannon's mathematical theory of communication offers a formally precise explanation of this fact, the informal slogan "garbage in, garbage out" allows us to grasp in a much more intuitive way what is at stake here: conclusions can only be reliable in terms of the data on which they are based. Thus, if the database is flawed, it will lead to biased and, sometimes, unfair and inequitable results. Therefore, epistemic and evaluative issues converge here.

Furthermore, if we assume that there is an inevitably normative burden on information technology in general and on algorithm development in particular (e.g., Newell and Marabelli, 2015), it appears that algorithms inexorably lead to biased decisions. The design and functionality of an algorithm reflect the values of the designer and their intended uses. As we have already anticipated, development is not neutral: there is no objectively correct choice in any instance of development, but many possible choices (Johnson, 2006). Consequently, it is difficult to detect latent biases in algorithms and the models they produce.

Friedman and Nissenbaum (1996) argue that biases can arise from at least three instances:

- 1** Pre-existing social values of social institutions, practices and attitudes from which technology arises.
- 2** Technical restrictions.
- 3** Emerging aspects of a usage context.

Social biases may be deliberately embedded in a system design by designers, for example in manual adjustments of search engine indexes and criteria to build rankings. However, they may also be unintentional, as, for example, in machine learning algorithms trained from human-labeled data which inadvertently learn to reflect those biases.

Technical biases arise from technological constraints, errors, or design decisions that favor particular groups with no underlying guideline values (Friedman and Nissenbaum, 1996). For example, when an alphabetical listing of hotel chains leads to increased sales for those listed first in the alphabet.

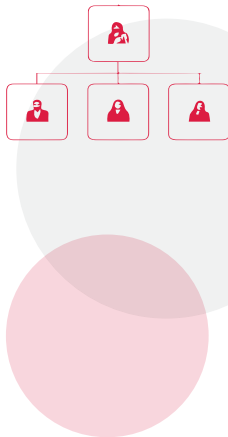
For their part, emerging biases are linked to advances in knowledge or deliberate changes in the uses of the system.



For example, clinical decision support systems (CDSS) are inevitably biased toward treatments included in their decision architecture.

The point we are interested in making here is that strongly discriminatory results can emerge from biased decision making and evidence. Indeed, profile-building algorithms (broadly defined as pattern construction or inference through data mining and the application of people's profiles whose data matches them) are frequently cited as sources of discrimination. These algorithms identify correlations and make predictions regarding the behavior of a certain group.

Bias can thus be conceived as a dimension of decision making itself, whereas discrimination would be a result of the decision in terms of a disproportionately adverse impact resulting from algorithmic decision making.



It is assumed that it is possible to direct algorithms in such a way that they do not consider sensitive attributes that contribute to discrimination, such as gender or ethnic group, based on the emergence of discrimination in a particular context. However, profiles constructed from presumptively neutral characteristics such as a simple postal code may inadvertently be juxtaposed with other profiles linked to gender, sexual preference, etc.

Romei and Ruggieri (2014) report four strategies to prevent discrimination in analysis:

- 1** Controlled distortion of training data.
- 2** Integration of anti-discrimination criteria in the classification algorithm.
- 3** Post-processing classification models.
- 4** Modifying predictions and decisions to be maintained

For their part, another practice, which is related to personalization, is also frequently discussed as a source of discrimination.

Indeed, personalization can segment a population so that only some groups receive certain opportunities or information, thus reinforcing pre-existing social disadvantages. Personalization through non-distributive profiles of the type seen, for example, in personalized insurance appraisal, can be discriminatory by violating both ethical and legal principles regarding the equal treatment of individuals.

From a philosophical point of view, John Rawls (1971) proposed that the criterion with which goods and services should be divided between men and women is justice. Rawls identifies justice with equality, a term he used with a political dimension since he believed that caring for the most disadvantaged was one of the State's obligations. Indeed, for the focus of *The Theory of Justice*, the title of Rawls' most famous work, the focus of the ethical analysis of our actions must be on justice as equality. The goal is to achieve a system of government that is able to have the appropriate political, social, and economic effects on society to guarantee this equality.



The simplest way to understand his proposal is through the mental experiment proposed by Rawls, a fiction that is useful to make his conception of justice more intelligible. Let us consider an original situation in which all men and women in a society agree on the way resources will be distributed, thereby leaving everyone satisfied. It would be an ideal order, which may never be achieved but to which we aim as the final objective. The point is that we must carry out this deliberation with a veil of ignorance, that is, making decisions without knowing what our identity is in terms of sex, gender, age, education, income, talents or skin color. Not knowing what I am like or under what conditions I will live in the world, I will determine what I think would be a fair and equitable distribution of wealth without personal biases or interests. By deliberating with our peers we will reach a reflective balance when, through opinions shared by the entire community, a consensus is reached around what is fair. In this way, for Rawls, justice as equality would guarantee the acceptance and tolerance of rights and freedoms for all because no one would know, until the veil of ignorance was lifted after agreeing on a just order, what their condition effectively is: if they are men, women, trans or non-binary people, rich or poor, whether they live with a disability, etc.

Once these principles are established, five steps must be taken to achieve a just society, namely: execute the contract, accept it unanimously, include basic and inalienable conditions (such as human rights), maximize the well-being of the most disadvantaged and ensure contract stability.


For the defenders of the Theory of Justice, this fiction serves to make it clear that freedom and equality are the manifestation of a democratic society based on free and fair cooperation between citizens, including respect for freedom and interest in reciprocity. Rawls believed that, in this way, his theoretical framework took the best of utilitarianism, since it sought to maximize happiness, but without its risks, since basic freedoms and rights were guaranteed while putting equality in first place as the central characteristic of all human social organization.

Along the lines of this philosophical reflection, we believe that a new option could be added to the list proposed by Romei and Ruggieri (2014) as a strategy to prevent discrimination: the extrapolation of the “veil of ignorance” mental experiment when designing AI.

Indeed, it would be enriching if at the time a certain algorithm is developed, the designers did it thinking of not knowing what their identity is in terms of sex, gender, age, education, income, talents or skin color. By not knowing what they are like or under what conditions they will live in the world, a fairer and more equitable distribution of opportunities could be favored (at least in principle) without the open impact of personal biases.

As we have pointed out, the development of artificial intelligence (AI) systems and their social application gives rise to ethical dilemmas and complex questions. By placing reflections in a hypothetical setting, we believe that the scrutiny of arduous philosophical questions is facilitated.

A hypothetical scenario:



Being aware of a conjuncture of economic and social crisis, the authorities of an Argentine province decide to launch a line of soft loans so that low-income people may have access to funds in order to alleviate their situation. As the demand is overwhelming, it is not possible for employees to analyze and evaluate the particular situation of each applicant, so the provincial bank submits tenders and purchases a model based on an automated learning algorithm that allows for financial scoring and credit analysis of all interested parties, yielding results that assist the person who will make the final decision.

Thus, those people who wish to have access to these loans must fill out their data on financial and non-financial competences on a digital platform, including their socioeconomic profile. As a result, an applicant report is created highlighting their weaknesses and strengths from the financial behavior analysis, the type of consumer and consumption profile. The system in question then offers recommendations on how to proceed with the loan application, measures the level of credit risk and compares it with commercial reports.

However, we have seen that based on the life cycle of machine learning, namely, training data, algorithm training, implementation in real situations, result assessment and parameter adjustments, there are a series of associated risks. In this section we will point out a crucial risk: the appearance of biases in decisions. Let's have a look at the possible sources of bias that we have considered

1. Pre-existing social values of social institutions,
2. Practices and attitudes from which technology arises.
Technical restrictions.
3. Emerging aspects of a usage context.

In our hypothetical scenario, factors of the first type could clearly have an impact, for example: pre-existing social values when considering the social context of the Buenos Aires conurbation (in Argentina). Feminization of vulnerability could be deepened there: given the fact that many women are heads of household and, at the same time, unregistered workers, the association of these two conditions in the system's training data could lead them to be excluded as subjects recommended for a loan application. In this way, inequality would be exacerbated instead of favoring the intended inclusion.

Likewise, in light of the other two types of factors (technical restrictions and emerging aspects of use), erroneous decisions could take place based on spurious correlations in the training data and in the subsequent creation of the applicant's profile. The possibility of discrimination based on the biases considered from this hypothetical scenario gives rise, then, to some questions that we would like to reflect on next.

Discussion

How can we avoid discrimination? For example, could a mere postal code exclude an applicant a priori at the input data level?

Could categorization be considered a form of indirect discrimination? Could different applicants/users require different treatment so that it is sensitive to their socioeconomic context?

As is well known, the system requires a vast collection of user data. What kind of sensitive information will be used and how it could be used in the future are questions that developers should make explicit to those who apply for a loan application.

Are these issues against which the State should guarantee protection? What kind of control should be exercised over data use to avoid discriminatory scenarios?

■ Privacy



The dividing line between private and public is difficult to draw. What data and experiences we want to share and what we do not want to share, and with whom we want to share them is a fundamental question that shapes our human life. This is a matter related to basic human rights: the right to privacy, the right to private property, the basic and elementary human right to draw a line between what is mine and what belongs to others. And there are also cultural variations on what counts as private and what is in the public domain. Let us remember the subversion in the order of which actions are public and which are private in the film *The Discreet Charm of the Bourgeoisie* by Buñuel (1972).

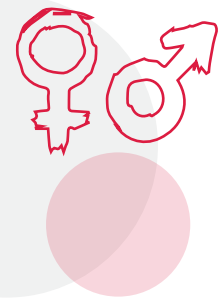
In the information age, in addition to worrying about taking care of the physical objects that are our private property (our land, our house, our family), we have to take care of the information we generate, our digital data. Intangible, difficult to grasp, difficult to care for. Partly because we are not aware that it exists, partly because we do not know how to take care of it.

Thus, we have a clear idea that our house is our possession, but we are not aware that our affiliation data (identity number, sex, age, name and surname, etc.) are ours and private, and that we have the right to share them with those we voluntarily want to, and that we can deny access to our data to whoever we want. Just as no one can compel me to let them enter my home, no one can compel me to tell him my name.

However, AI developments based on machine learning are only possible because they process a huge amount of data and "learn" from it, largely personal data from thousands or millions of individuals. And even if that data is anonymized, it is still personal data. This fact has generated one of the biggest controversies in AI ethics.

Although there are various laws in force regarding people's digital data, the complexity of this matter motivates at least two philosophically relevant questions, which could give rise to different problems regarding data.

Firstly, the issue of legally determining (in each country) which data is considered private property belonging to individuals and which is public is still pending resolution. The data considered to be private should be in the possession of each human being, in such a way that each of us is the only person who is in possession of that data and is not forced to share it. Note that our information regarding identification is usually information that we are forced to share constantly: migration forms, bank, tax, school, and medical forms, etc. inescapably include personal data. Although it is usually a little discussed topic, some doubts have recently arisen about the need to share certain data. Thus, feminist movements in favor of sexual diversity question the need to include the feminine/masculine binary option in many of these forms. Note that this cannot be solved by simply adding more options. The problem is even deeper. Why should we make our sex public when we are applying for a bank loan, for example? If one thinks of the matter in these terms, the answer should certainly be that it is absolutely irrelevant what sex we have in this case. On the contrary, it is highly relevant in medical contexts where the female/male pair falls short: identifying a trans woman as a woman or a trans man as a man could obscure relevant medical data in certain contexts. Personal data is what constitutes our identity, our sexual, racial, class, and professional identity, etc. Without a doubt, it should be up to each of us to decide with whom and to what extent we want to share our personal identity.



Secondly, there is the problem of data usage. Every time we fill out a form (be it in the already mentioned cases or be it to access some Internet site, app, etc.) we are making our personal data public. Now, how public are we allowing that data to be? There are contexts where we transfer data knowing that it is confidential, for example in the doctor-patient or lawyer-client context. But in most other cases it is not clear how public the data we provide when filling out forms can become. It is also unclear whether we "loan" data to those who ask for it and we are able to recover it, and whether the data can be "returned" to us whenever we want. Because meanwhile, that data is being used for many things without our consent. Or perhaps with our consent, since we have clicked on "yes, I accept" after (not) having read the "bases and conditions" that, in general, are written in a way that is difficult to understand and are very lengthy to read in detail.

On the other hand, the mere fact of our existence in the world today generates "data" that is accessible to computer systems through security cameras, microphones, photos of ourselves that other people

take and upload without our authorization to Internet websites (no Internet website requests authorization from all those people who appear in photos and videos so that their image or voice is made public).

Thus, the document "For a trustworthy AI" of the European Union ("Building trust in human-centric artificial intelligence") recommends protecting personal data in all instances of development and use of AI systems. In order to achieve this, the first step in any design should be to anonymize personal data in such a way that it becomes unidentifiable with the flesh and blood person who entered it into the system so as to use that data to generate machine learning systems. In this way, the same data would be useful to generate AI systems, but the person who submitted their data would not be in danger or at a disadvantage. For example, a system could be generated to help diabetics regulate their daily insulin doses, without being able to identify which individuals with diabetes submitted or which ones did not submit their data to create the app in question, i.e. the learning algorithm was run to generate the model based on data from these people.

However, this proposal has limitations because many of the systems that are generated continue to feed on personal data in order to reach concrete flesh and blood people with their suggestions. What would be the use of a general system that helps people take care of their insulin level if we do not give them specific data on a specific patient to see when that patient needs an injection? Non-anonymous personal data is essential for systems to reach specific people (with suggestions for films, medical treatments, etc.).

The original proposal is still pending. What personal data can be used by whom and what for? The discussion remains open. Can a private company use personal data to offer specific advertising to each user? Can a government (democratic or not) take personal data from people on surveillance cameras? For what purposes would said government be authorized to do so? To apprehend criminals with an arrest warrant, or to arrest suspects as well? (What is considered suspicious - a person's actions, facial features, skin color? What data was used to train this system for assessing "suspects"? And so we return to the question of biases.) Would the government be authorized to identify leaders of protest marches against a government? Is it possible for a prospective employer to have access to someone's medical record data to make decisions about whether or not to employ them? Is a scientist able to take data from medical records to do research and publish an article on a particular topic? And what about the thousands of data generated in imaging studies from medical contexts? And, in particular, in the case of brain data, who can use it and what for?



Recently, in the journal *Nature* (Yuste et al., 2017), members of the BRAIN project (a medical-scientific and technological initiative aiming to technologically reproduce the characteristics of the human brain) have drawn attention to the need to legislate on brain data, proposing said data to be considered as a basic human right: each human subject owns their brain data because our brain data is our soul and our identity as human subjects. This would mean leaving decisions about the use of these data to doctors. No one could use personal data for commercial purposes, or sell or buy data, in the same way that it is forbidden to trade, sell or buy other human organs, such as kidneys for example.



Nowadays, brain data is used by researchers in medical sciences, for example, by means of automated learning algorithms on thousands of brain imaging data from patients who later developed some neurodegenerative disease such as Alzheimer's, in order to see brain evidence of the disease development before it is detected symptomatically. Undoubtedly, it is beneficial for the population to have an early detection system for Alzheimer's patients to promote preventive treatments, but the cost is the use of very personal data from thousands of people who have not necessarily consented to provide it. And these developments and data can also be used for other purposes, for example to generate brain-artifact designer interfaces for various purposes, such as orthopedic hands, but also super-soldiers, i.e., weapons of mass destruction.

In short, people's digital data must be safeguarded to avoid situations of injustice, bias or disadvantage by those who handle our data and their actions towards us. But it is also true that personal data that is not voluntarily submitted is generating systems designed for the benefit of humanity, such as the case of Alzheimer's or systems that allow the police to locate fugitives by using face recognition systems on security cameras on public roads. The complex dividing line between private and public and the potential misuses of personal data should be a constant concern for AI designers and users.

Let's consider an imaginary case

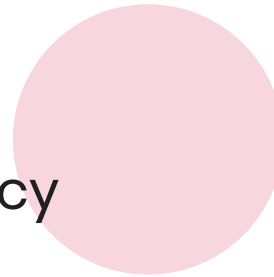


The brand new A.L.F.R.E.D. voice assistant lets you control various home devices - like music, lights, and the washing machine - as well as make grocery lists, send text messages, and answer questions about the weather and traffic. To activate it, each user must say "A.L.F.R.E.D." and the required action. In order to stay alert, A.L.F.R.E.D. is constantly detecting sounds in the house, but its programmers made sure to maintain a rigid principle of privacy: it is impossible for the voice assistant to have access to the recordings of what it hears or to react if it is not explicitly called. Now, what happens in the case of a house where domestic violence occurs? What if, when listening to certain typical phrases of dangerous situations, the assistant activates itself by asking for police assistance or by trying to dissuade the attacker? What is worth more: a person's privacy or their integrity?

As AI designers, we should ask ourselves what right we have to incorporate a detection system for dangerous situations such as those described into a device like A.L.F.R.E.D. Who authorizes us to incorporate this function? One option is to leave it up to the user when configuring A.L.F.R.E.D. to leave this function active or not. Of course, the person who configures the device is probably the person who exercises violence... and either deceives himself and does not think he is exercising violence, or does not want this situation to be recognized and probably decides to configure it in such a way that it is deactivated. Is it possible for the manufacturer to include this function without user consent? Now let's think about the situation on the side of the person receiving the emergency call, i.e., 911. Is the system ready to receive automatic calls of this kind? Does the system have to respond differently in these cases? Why?



■ Transparency



The transparency of a technological system is not independent of the existing technical culture in a society. A strong, sophisticated and widespread technical culture in a society allows its members to participate in an informed and conscious manner of the public debate regarding what technology should be developed, for what purposes, in what environments and which point of view it has on human life. A technical culture incorporated in a society generates a background of understanding of technological developments that eliminates prejudices and allows the pros and cons to be prudently balanced according to the way in which that society wishes to live or the way in which it understands what a life worth living for is all about. A proper ethical assessment of artificial intelligence systems requires effective public communication of artificial intelligence characteristics and its impact on the life of societies.

Technical culture is understood as the set of knowledge and representations about artificial objects, their components, structures and functions, in addition to practical components such as rules, skills and operational knowledge regarding artifacts and their functions. Finally, technical culture also includes evaluative components involving the objectives and results of the functions of the devices and of our actions with them.

Understanding artifact or technological system designs depends on the existing technical culture in different social groups. The presence of adequate technical cultural content is the hinge that connects the user with the designer and what places the artifact in the context of its cognitive, technical and cultural history. In this way, potential users manage to understand the purpose of the artifacts, use the appropriate conceptual models to understand their possibilities and restrictions,

adequately evaluate the results of their functions and develop and modulate trust/mistrust relationships with artificial, emotional systems, etc.

It is an ethical requirement of artificial intelligence technology to communicate and disseminate its features and applications so that citizens can make informed decisions. The transparency or explicability of AI systems is directly associated with the autonomy of human life. If there is no minimum understanding of the technological systems that incorporate AI, the exercise of autonomy cannot be guaranteed since citizens must always surrender their authority to the experts and will not be able to judge lucidly by themselves the adoption or not of that kind of technology. Without an understanding of technology, there is no judicious and adequate evaluation of its reality.

The primary components of transparency are: information **(1)** accessibility and **(2)** comprehensibility (Glenn and Monteith, 2014; Kitchin, 2016).



(1) Information about the way in which algorithms work is often kept secret for reasons of competitive advantage, ownership or national security. In these cases, it is argued that transparency could counteract the autonomy of organizations and data privacy. This generates an asymmetry in information and an imbalance in the knowledge and decision-making power in favor of data processors (Tene and Polonetsky 2013).

(2) Beyond the knowledge of the data and algorithms on which a system is based, the information that the system provides the user as process output must be understandable by the user in order to be considered transparent (Turilli and Floridi, 2009). Efforts to make algorithms transparent face an arduous challenge. The persistent problem of interpretability in machine learning algorithms points to the challenge of opacity. Indeed, the algorithm modifies its behavior structure during the operation (Markowetz et al., 2014). This alteration about the way in which the algorithm classifies new inputs is precisely what allows it to learn (Burrell, 2016). The underlying reason (rationale) for the algorithm is obscure, giving rise to the so-called "black box" problem. The algorithms are opaque in the sense that the receiver of the algorithm's output has no idea how or why a particular classification was arrived at based on the inputs. Both the inputs (data) and the outputs (classifications) can be unknown and even unknowable. Opacity in machine learning algorithms is a product of the high dimensionality of data, of the complex code, and of the changing

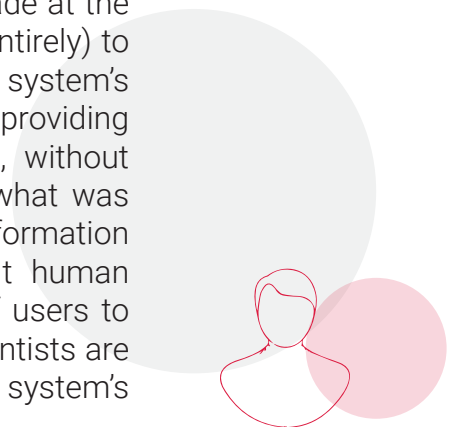
decision-making logic (Burrell, 2016). Likewise, meaningful consent to data processing is not feasible when opacity makes risk assessment impossible (Schermer, 2011).

Requests for transparency often point to the possession of open access codes or acceptance by the user of the terms and conditions of a service. However, it could be said that there is a richer notion of transparency that involves sharing the ends, the means and the thought processes behind engineering decisions with those who are affected by them.


Along these lines, it is important to note that transparency could be a demand about the **(I) algorithm** or about the **(II) use** of the system in question.

As regards **(I)**, considering the limitations of the “black box” explained above, developers could question the plausibility of being transparent with machine learning algorithms. They sometimes contend that they themselves could not decipher or reverse engineer the way in which specific algorithms operate after having trained them on vast data sets. The purported value of these algorithms lies in their ability to solve decision making in a way that would be superior to - while optimizing - the possibilities of human processing. Algorithmic transparency would therefore not genuinely contribute to user understanding.

Let's consider then what happens with respect to **(II)** transparency about use. In this case, the aim is to make the decisions made at the time of system design sufficiently understandable (but not entirely) to the user so that the user can develop confidence in the system's reliability and underlying justice. In general, this involves providing some information about how the system was developed, without technicalities, and according to the technical culture (see what was said earlier) of the user. Given that the degree and type of information that different users can assimilate is variable in different human groups, transparency is relative, it varies from one group of users to another and experts in both computer design and social scientists are constantly required to perform the task of “translating” the system's technical specification for effective users.



Let's consider some of the issues that this situation raises from the consideration of a hypothetical scenario.



One of the main obstacles in the fight to eradicate Chagas disease, a disease caused by the parasite *Trypanosoma cruzi* and which represents one of the greatest health problems in rural regions of Latin America, is that the population in danger of contagion lives in regions with little access to the public health system or in remote places. Faced with this situation, the HealthCare company developed a system that ensures detection with 85% certainty for the presence of triatomines, which are responsible for the disease, in the person's environment just by analyzing images of a person's saliva. So the company developed an expensive application for cell phones that uses the device's high-definition camera to take pictures and send them to a cloud server, where they are compared by automated learning algorithms with other similar images to determine whether or not it detects elements compatible with triatomines. A Latin American country buys 10,000 licenses for this computer solution and distributes them by means of brochures in markets and bars to the rural population that could be ill. Although the application requires accepting its terms and conditions to be used, in addition to the warning about its accuracy reaching only 85% of cases, many users cannot read or do not have their own cell phone, and use a borrowed one. In addition, the objective of the app is that the results obtained are interpreted by a doctor in order to provide a diagnosis, but there are no professionals available and many people replace their analyses with what the app indicates.

There are many issues that will be taken into account during the development of a system such as this, for example, the material resources available, the economic, legal, and environmental feasibility, the algorithm programming and the database determination of the cases that will be fed to the system and from which it will learn (whenever a technology is developed with limited data from a

population and the results are extrapolated to another population, the possibility of extrapolating development without change to a new human group should be checked). But let's assume that the developed system is robust and reliable.

In this example, it is especially interesting what happens with the transparency of the technological system compared to the fact that the user does not know how the algorithm operates, despite the fact that it provides correct results in a high percentage. The adoption of this type of technological development by a population whose technological culture is limited generates complex situations. If the population is not familiar with the use of mobile phones, how will they understand the diagnostic possibility of the device? What will the user do with the information received? A diagnosis without a subsequent treatment is useless, and only produces anguish in the person who receives it. Does it make sense to distribute this development directly to the population or does it make more sense for it to be available in health centers so that there is a health professional in charge of its use?

In the latter case, the doctor-patient relationship becomes central, a value niche in which the technological system is inserted. And beyond the fact that the organic diseases of human beings are the same throughout the planet (although with a different incidence in different populations), the ways of receiving and dealing with the information in question may vary from culture to culture. Therefore, if the system is being designed to be used directly by physicians versus patients, it should be coupled to interfaces that take into account the particularities of the population and the way in which medical care and information transmission are provided in the doctor-patient relationship in that culture. The adaptation of the technological system to the sociocultural niche in which it is inserted is a basic ethical demand on technology development. In this case, development transparency means making relevant information about the way in which the system operates accessible to health professionals so that they can trust the system when making diagnoses in the population under treatment.

As can be seen, the transparency issue is relative to the user's technological culture for whom the development is intended, so that making the system operation "transparent" requires specialists in human and social sciences who know the public to whom the explanation provided of the system operation is intended.



■ Responsibility

As it is clear from previous sections, the implementation of new technology implies, in addition to promises of change and transformation, risks for both individuals and society in general. In the case of the current scenario of AI models, we can see how these developments occupy more and more space in critical decision contexts such as the health industry, the judicial sphere, government agencies, financial organizations and warfare territories, among others. Faced with this, it is imperative to start a discussion on who would be responsible for the consequences and impacts of these advances.



Some possible conflict scenarios are:

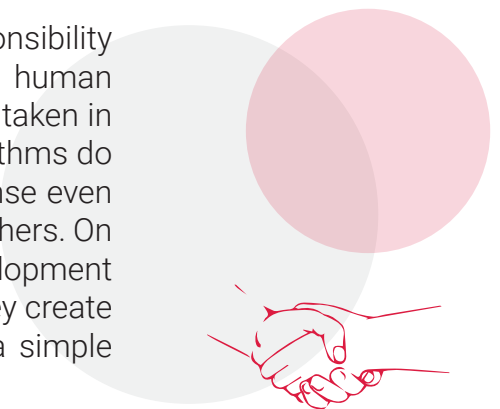
- High-tech devices such as autonomous passenger transport or surveillance vehicles have, for example, multiple components rarely provided by the same manufacturer that work together, such as sensors, cameras, processors, different types of internal software codes, and multiple automated learning algorithms. To what extent is each person in charge of a complex team responsible for the conduct of the device if these elements must be completely interconnected for the correct operation of said device?
- On many occasions, artificial intelligence developments are devised and developed in one country but used in various parts of the world, in communities that do not always share the same cultural and community values. Are their developers responsible if principles from other communities are violated? Is it lawful to prohibit the use of certain developments in certain territories?

- Who is responsible if an algorithm is biased? Those people who programmed it, those who provided the database, or the company that uses it to its advantage?
- The people who develop technology that will replace jobs done today by men and women, in what sense are they responsible for changes in the working world and in the disappearance of those jobs?
- If it is true that many deep learning systems function as black boxes and that the mechanisms leading to their conclusions are not transparent and cannot be audited by human beings, then who is responsible for them?

We are, then, faced with dilemmas and problems about responsibility that seem to be particular to this area. For example, while human agents can be called to account for the decisions and actions taken in the event they have affected others, AI models and their algorithms do not appear to be responsible in the same morally relevant sense even though they make decisions and take actions that can affect others. On the other hand, the agents involved in the creation, design, development and implementation of these models are so numerous that they create a complex network where establishing responsibility is not a simple matter.

It is not a minor fact either that the attitude of men and women towards artificial intelligence devices tends to be different from that of other devices. Sometimes they are perceived as superior and more advanced in their abilities, generating a false sense of confidence. But there are other cases where unfounded fears and prejudices raise suspicions about the results of this kind of progress, potentially depriving us of the benefits that could be produced. One of the necessary steps to achieve true confidence in artificial intelligence is to clearly determine the roles and responsibilities of each player involved in its development and application, which includes not only companies and States but also developers and users, among others.

The World Wide Web Foundation, for its part, distinguishes algorithmic responsibility - that is, the obligation of algorithm designers to clarify what the potential harm of their developments is - from algorithmic justice,



which is related to the ability to repair damages (Webfoundation.org). This differentiation helps shed light on the need to think about liability before any harm occurs and how those people who think, design, and develop algorithms must also anticipate the undesirable consequences of their work.

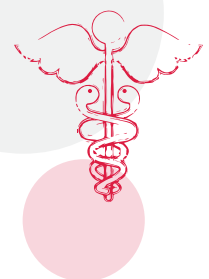
The question of responsibility, then, does not appear when the system starts up but at its very genesis. The moment designers decide that a given design is the best solution to the problem they are trying to solve, they become the agents responsible for the modifications introduced into the environment, since the effective realization of a design changes existing reality and people's lives, regardless of whether or not the consequences that these changes bring are perceived in advance. Thus, there is no design without responsibility for what is designed, which concerns the people who imagined the artifact or the technological system, transforming it into a real event in the human environment.

Another important point, beyond the robustness of the technological system, is the responsibility for system failures. This raises a previous question: should the expert suspend their judgment and defer it to the system? Who is in control of the situation - the medical expert or the system? Responsibility is not independent of the placement of the ultimate source of authority. Basic ethical requirements require that responsibility for false positives always rest with human authority. Furthermore, the system must always be under human action and supervision. Results can never replace the best judgment of the human expert. Ultimately, the expert's judgment is used to calibrate the diagnosis issued by the machine; otherwise, no one could be held responsible for the results.

In the current literature, we can distinguish three senses in which we speak of responsibility related to artificial intelligence:

1. Responsibility as a feature or trait of the artificial intelligence system itself.
2. Responsibility in the narrow sense regarding individuals or groups that are accountable for the effects of technology.
3. Responsibility in the broad sense that corresponds to the socio-political-economic ecosystem that develops, acquires, implements and uses artificial intelligence advances.

The first of the senses is linked to the need to incorporate explicability into artificial intelligence systems and algorithms, while the second and third focus on being able to determine who, in the wide ecosystem of agents involved, is responsible for or the most responsible for the effects produced in a specific event or in society in general.

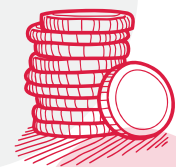


And although transparency seems to be one of the necessary conditions when thinking about the responsibility of artificial intelligence – insofar as a transparent algorithm allows, for example, a better scrutiny of its internal mechanisms-, this feature does not seem to be enough since transparency alone is not enough to better determine the different parties' responsibility. There are other principles that must be present, such as explicability, the capacity for responsiveness, and auditability.

A responsible AI project should be responsive as regards being able to justify decisions that emerged algorithmically with human support. This means that an end-to-end chain of responsibilities must be established with people who take responsibility for each step taken, from design to implementation, in order to understand the results provided by the algorithm. It is about being able to offer explanations of the production processes to any individual who request them in a clear, understandable, coherent and accessible language for non-technical people.

As regards auditability, it is a necessary condition to be able to determine both responsibility for design and usage practices and justification of results. In order to do this, developers and implementers of algorithmic systems must keep records at all stages, from the first developments, of the modeling of test versions, of data collection, etc. These records should be accessible to peers and supervisors as a necessary, but not sufficient, condition for the procedures and operations of the AI model to be safe, ethical and fair.

The global economy of applications and technological developments also pose challenges when evaluating responsibilities, since it seems to be increasingly difficult to think about these advances within certain borders. Each algorithm has in its conception the set of cultural assumptions of the people who designed them but, on many occasions, they are used or implemented in foreign contexts where there are other norms or principles. This also generates not only cultural or idiosyncratic but also legal and regulatory clashes. The discussions about privacy, for example, make it clear that these problems occur in the clash of jurisdictions and that their solution requires the participation of multiple stakeholders and multilateral coordination.



Some possible conflict scenarios are:

The complex network of agents involved in the design, development, creation and implementation of algorithms imposes a dialogue of the many agents involved, since it is a terrain marked by complexity and intersectionality.

A non-exhaustive list of stakeholders should include:

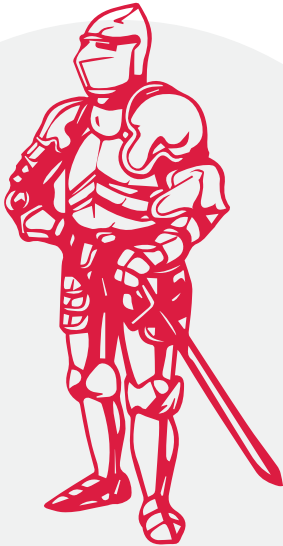
- State representatives, legislators and public policy makers, who must promote responsible and inclusive behavior.
- Intergovernmental organizations, which must provide a framework for discussions by all stakeholders across borders.
- Private companies linked to the world of technology, which have to respect codes of ethics when handling sensitive materials such as data, and whose products need to be auditable in addition to complying with local legislation.
- Members of the academic world, who on the one hand must guarantee that the ethics of artificial intelligence is present in the curricula of the careers involved while simultaneously researching and offering tools for reflection of these practices, facilitating in this way interdisciplinary research.
- Non-governmental organizations interested in designing mechanisms for inclusive citizen participation, being bridges of knowledge exchange and dialogue between the public and private sectors.

Let's imagine the following case:

After years of patient development and testing, the Hefestus company has "HER-CULES" ready, which is a state-of-the-art exoskeleton that allows anyone who uses it to effortlessly move large and heavy objects. It is a kind of pneumatic super-suit with sensors that, when detecting that a movement could hurt someone, automatically cancels any action.

The port company Container S.A. acquires six units of "HERCULES" and trains a dozen employees in their use, increasing their transport and logistics capacity. Two years later, Hefestus detects a fault in the sensor system and sends a firmware update to fix this problem. However, due to the economic crisis that the country is going through, Container has stopped updating exoskeletons because each update comes at a cost in foreign currency.

Weeks later, an operator seriously injures a colleague while working: who is responsible? The "HERCULES" designers, who did not foresee the failure? The Hefestus company, whose business model involves paid training and updating to use its products? The Container authorities who stopped acquiring the improvements? The employee who was using the equipment?



■ Security

We have pointed out in this work that the impact of implementing AI models is vast and complex. It has consequences in individuals and in society due to its power to change or consolidate inequities, because it can repeat biases and prejudices already installed, because it has the potential to put our privacy at risk and because, if there is no concern for liability and transparency, the consequences can be astonishing without anyone taking responsibility for it. But there is another dimension that must also be taken into account, and that is security. A good AI model must be reliable, accurate, robust and safe.

Ulirresponsible data management, a negligent design or unreliable production and implementation processes can each lead in its own way to dangerous or unsafe outcomes that directly or indirectly harm people's well-being and public well-being. They can also undermine society's confidence in the responsible use of AI technologies.

It is a great challenge because these systems are implemented in a reality full of unexpected events, with great uncertainty and volatility and where it is possible for crimes to be committed or for the systems to be used for a purpose other than the one intended, for example to cause harm. In order to be truly useful, AI models must be flexible enough to be used in real scenarios. However, this is where one of the main problems for security lies: without enough controls, a model can determine a course of action optimizing resources and achieving the best result but by means that are harmful to people. These models do not always have the ability to read different contexts and lack some of the tools that men and women have to deal with the unexpected, such as common sense or empathy, so they can have harmful consequences that their programmers did not anticipate or limit. This is an arduous problem, since the limitless complexity of the world makes anticipation of all its circumstances and variables seem impossible.

This reinforces the need to worry about producing secure AI technologies, with mechanisms to mitigate risks of failure when facing the real-world scenario and its unforeseen events, as well as minimizing the possibility of being vandalized or becoming harmful tools that produce detrimental results which in turn undermine public confidence. Building an AI system that meets these goals requires rigorous testing, validation, and reevaluation, as well as integrated adequate monitoring and control mechanisms for its real-world operation. In turn, the consequences of lack of security vary significantly according to each model, since an algorithm that distinguishes between valid and spam emails is not the same as an algorithm that recognizes potential targets in a military defense system.



The reliability of an AI model occurs when it behaves exactly as its creators and designers intended and anticipated by meeting the specifications for which it was programmed. Reliability is, in this sense, a kind of consistency measure. However, model users do not always use them with their intended functionality, so this original objective is betrayed. That is why, beyond the warnings that can be made - with terms and conditions of use, in manuals or in training, for example - it is necessary to take into account the way in which systems can be used erroneously, whether intentionally or not, so as to minimize any ability to do harm.

Being able to establish the security of a system also depends on the precision of its behavior. Although this measure changes according to the characteristics of each AI model, we can define the precision of a model as the proportion of examples for which it generates a correct result. The choice of a certain level of precision or an acceptable error rate is the evaluation measure of its performance, and this will be what determines whether it works well or poorly. A malfunction is the first sign that system security is compromised or that there are external attempts to damage it, and besides, the level of trust that will be placed in said system depends on how many errors it makes. In a volatile and changing reality, in a scenario full of uncertainties, it is not always easy to ensure precision.

In these unpredictable conditions, there is a need for AI systems to be robust, that is, they should operate reliably and precisely under adverse conditions. These conditions may include harmful external interventions with the intention to commit vandalism or radical and unexpected changes in any of the factors that do not normally change or that can be found in the expected standards. The measure of robustness is the strength of a system and the solidity of its operation in response to difficult conditions or attacks.

Thus, in addition to having a correct performance, the objective of an AI system's security is to be able to provide protection against possible attacks or malicious actions, always maintaining the integrity of its information. The confidentiality and privacy of their own and user data must be safeguarded at all times. This implies, on the one hand, ensuring that only authorized people will have access to use it and that there are mechanisms to protect its internal architecture from any unauthorized modification. Its operation must also be foreseen despite damage to any of its components or under hostile or adverse conditions, such as in the midst of an external attack.

Adverse or malicious scenarios are numerous and it is not possible to provide an exhaustive list, but we can point out certain conditions that must be foreseen when designing AI models. On the one hand, most machine learning systems are built with a static image of the reality in which they will be operating, which emerges from the historical data with which they have been fed and which defined their internal parameters. Compared to this still photograph, when these systems are applied "to the real world", their precision and reliability are especially vulnerable to changes in the distribution or composition of this data. When the historical data that has been crystallized in the trained model architecture ceases to reflect the relevant population to which the model is applied, it quickly renders the system error-prone in unexpected and potentially damaging ways. Staying on top of these data transformations is crucial for a secure AI, and each team must ensure an action plan to anticipate and reduce their eventual impact.

The incorporation of new or unexpected information can undermine the robustness of any system that, like deep learning systems, bases its performance on the processing of massive amounts of data to draw from, resulting in thousands or even millions of parameters. These models may have difficulty in processing unknown events and scenarios, making serious and unexpected errors because they do not have the ability to contextualize problems they are not programmed to solve. As we have seen in the transparency section, these errors can remain inexplicable given the high computational complexity of their mathematical structures. This fragility can have very adverse consequences when systems are applied in autonomous vehicles, such as cars or military drones.





Malicious behavior specifically intended to make an AI model malfunction must also be considered. There are several well-known strategies, but as these models become increasingly popular in state, business, and personal arenas, the appeal for those seeking to profit illegally increases, which is why these attacks are becoming more frequent and different. One of the most common forms of damage is the intentional modification of input data, often imperceptibly, to induce the system to commit a misclassification or incorrect predictions. This is what happens, for example, when the pixels of an image are altered so that a face or a building facade cannot be recognized. The use of incorrect or damaged data can occur even before the system is put into operation. "Data poisoning" is a type of attack in which part of the data set on which a model will be trained, validated and tested is modified or manipulated. By altering a selected subset of training inputs, the AI system will misclassify, underperform, or have an "Achilles heel" so that once it is installed and supposedly working properly, an attack can occur by exploiting that vulnerability.

Some possible conflict scenarios are:

Some possible strategies to mitigate the risks of an AI model include:

- Run extensive and varied simulations of situations and contexts during the testing stage, so that appropriate constraint measures can be programmed into the system.
- Create transparent and auditable systems that allow continuous inspection and monitoring.
- Always include mechanisms that allow authorized persons to cancel or turn off functions and systems remotely and at any time.

Let's imagine this case:

In order to design advertising ads that are really appealing to social network users, who sometimes claim to be fed up with the number of ads they see on their feeds and timelines, the RelevANT company designed a machine learning algorithm model that takes all the public information associated with a user to create ads they find incredibly appealing. It is such a successful development that several companies bought it and reported an increase in their sales.

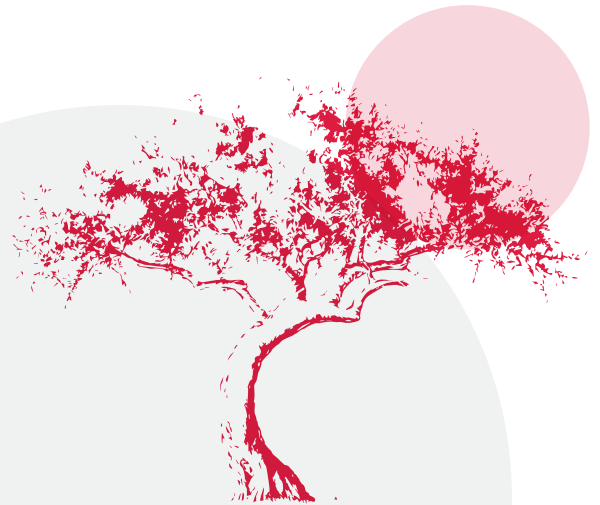
Aware of its effectiveness, a group of computer criminals bought the algorithm and used it for a phishing campaign, with which they flooded social networks with misleading ads that led users to hand over their credit card details or passwords to their online bank accounts, which ended up being one of the most profitable computer crimes in history.

What responsibility would you say RelevANT has for the crimes committed by their clients? Is it possible to limit the use of technologies as precise as those of an algorithm that creates a perfectly personalized advertisement for each individual? Is that desirable?

If the algorithm for personalized ads were contracted by a brand manufacturing a highly sugary soda, would it be reasonable to require the company not to use them with minors or populations with health problems?



■ Well-being, human rights, politics and technology



Well-being

Human beings pursue well-being, that is, we are interested in living but especially in living well (Ortega y Gasset, 1998). Technologies, which transform the world according to our values and purposes, are one of the main forces forging well-being. They incorporate the way in which human beings think of themselves and of the world. They portray their fears and desires and, as real solutions to problems related to the management of daily individual and social life, contain in fact different answers to the question of what good life is all about - for example, the practices, techniques and beliefs of the Quom peoples of the Gran Chaco region shape a way of answering the question of what well-being is which is completely different from the technological practices and beliefs that organize life for the Silicon Valley tribes (Sadin, 2018).

Well-being or good living is an always mobile term, unlimitedly variable throughout human history since men and women have not wanted, nor want, nor will always want the same thing; the profile of our idea of a good life is in permanent transformation and is adjusted based on our experiences, knowledge, desires, values, fears, fantasies, concerns, etc.

Well-being is increasing more and more– with no limits? - our capacities.

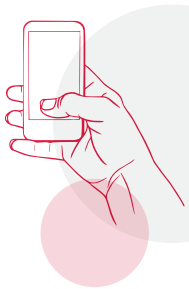
In order to design advertising ads that are really appealing to social network users, who sometimes claim to be fed up with the number of ads they see on their feeds and timelines, the RelevANT company designed a machine learning algorithm model that takes all the public information associated with a user to create ads they find incredibly appealing. It is such a successful development that several companies bought it and reported an increase in their sales.

Aware of its effectiveness, a group of computer criminals bought the algorithm and used it for a phishing campaign, with which they flooded social networks with misleading ads that led users to hand over their credit card details or passwords to their online bank accounts, which ended up being one of the most profitable computer crimes in history.

People deliberate to resolve issues related to the way we live and our idea of well-being. And we do it with the purpose of obtaining that intimate comfort of feeling at home with ourselves and with others. What guides our deliberations is what we consider important to us (Frankfurt, 2004), the reason for our care and attention. However, the ingredients of well-being are not so obvious. What well-being consists of is a fairly spatially and temporally contextualized issue. Different people, human groups and cultures advocate different contents for the notion of well-being.

The notion of well-being has an individual and a collective dimension. In its individual dimension, well-being is directly related to the use of technologies and how they contribute or not to improving people's daily lives and the realization of their life plans. In its collective dimension, well-being is related to human rights as transversal contents to contemporary cultures and societies. In this dimension, the fulfilment of human rights is a fundamental concern in all areas of human practice, including the design and implementation of artificial intelligence systems.

Individual and social well-being



Technology nowadays has a direct psychological impact on people's lives. A quick look at our daily habits shows that 80% of smartphone users look at the screen of their cell phone when they wake up, even before brushing their teeth. It is estimated that the average user spends at least 4 hours per day in front of their cell phone or tablet. Most of that time is spent on some social network. In fact, there is talk of cell phone "addiction", with many tips and treatments available online to treat it even though it has not yet been classified within the mental disorder framework. Although the results from these studies are diverse, they are consistent with the existence of a negative impact on users' quality of life, who spend a significant amount of their time in this digital world.

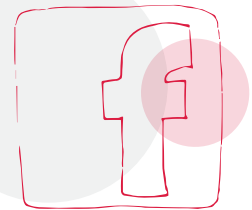
New problems related to the manipulation and control of our human evolutionary instincts also arise with new technology. Technology platforms exploit human experience, allowing addictions, undermining social relationships, spreading propaganda, manipulating children, etc. It is a massive, growing global problem that affects more than two billion people.

There are at least four categories of problems that directly affect individual and interpersonal well-being.

1. **Attention.** Constant interruptions from our technological devices undermine our ability to focus both on a task that requires attention and on our interpersonal relationships in everyday life - even when the device is turned off, it reduces our cognitive capacity (Ward et al., 2017).
2. **Mental health.** The average number of friends that people have has dropped in recent years. Although social networks have been designed precisely to connect more people, it seems that its effect has been rather the opposite. A strong correlation has been found between feelings of loneliness, depression, stress and the use of social networks (Zoller, 2019).
3. **Interpersonal relationships.** Having a conversation and a person-to-person contact is very different from having it through a platform. The latter create less emotional connection and run the risk of being constantly misinterpreted. We know there are many signals (both conscious and unconscious) that get lost in a merely digital contact (Turkle, 2015).

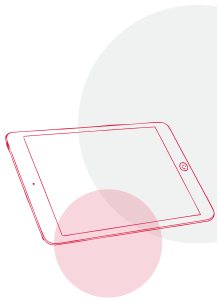
4. **Vulnerable social groups.** Children are a particularly vulnerable group to the interaction policies of platforms, particularly with regard to bullying and the use of their image. Although all platforms set an age limit so that children under the age of 13 do not enter, there is no effective barrier nowadays to enforce it. Moreover, the use of cell phones in school contexts significantly impairs the attention span necessary for the tasks required in children's education.

Current social media technologies try to maximize their benefits by trying to increase the amount of time the user spends on them (and thus be able to show them more advertising and earn more money). If this is the prevailing interest, then the goal of its design is on the opposite side of an increase in the user's quality of life. Should social media have warnings about the harm that may be caused by excessive use like the warnings implemented for cigarettes, alcoholic or sugary beverages? Would a simple notice as such be effective? Perhaps the most effective approach would be to think from the very design of the platforms, so that they are controlled by a regulatory framework that prioritizes the health of the general population. But what form should such a design and regulatory framework take? In principle, two aspects should be taken into account:



- (a) Problems must be made visible, stressing the fact that the use (and abuse) of platforms is not harmless at all. Both the State and the companies themselves can devise campaigns (together and in parallel) to limit their use and help those who have a problematic relationship with platforms. For example, the design of the platform itself must warn of the abuse that the user makes of them, reporting the time they spend on them and effectively showing the damage caused to them.
- (b) Special focus should be given to vulnerable users such as children. Sanctions could be implemented for companies that do not actively try both to unsubscribe active users recognized as minors and to forbid minors who try to subscribe.

Could companies be required to put a limit on possible addiction to their products, as has happened with the case of tobacco companies and cigarettes? Does the way social media is designed harm our psychological health and our social skills for leading a good life?



Another important aspect of new technologies, which is related to a person's well-being, is that they should be reversible, that is, any user must be in a position to return to an earlier state while using the technology if they consider that said technology is doing damage of any kind. Technological designs have to promote a non-alienating relationship with technology. This guarantees that the person can have control of the devices and that their autonomy prevails over the different forms of dependency that spoil the exercise of their freedom.

Well-being, human rights and democracy.

The creation and assessment of artificial intelligence should not be indifferent to showing respect and promotion of human rights, which are rights that are inherent to all human beings, without distinction of race, sex, nationality, ethnic origin, language, religion or any other condition.

The Universal Declaration of Human Rights includes civil and political rights (for example, the right to freedom of thought, conscience and religion; participation in public affairs and elections; protection of the rights of minorities, among others) and economic, social and cultural rights (for example, the right to work in fair and favorable conditions; the right to education and to enjoy the benefits derived from cultural freedom and scientific progress, among others).

Artificial intelligence can promote a loss of civil, economic, social and cultural rights or broaden its enjoyment. The key will lie in the way those technological systems are designed and implemented; in short, in the idea of well-being promoted by companies and the States behind the development of artificial intelligence systems, as well as the role that citizens adopt in the

face of challenges posed by this technology. The impact of artificial intelligence on human rights has an individual and a collective dimension. The individual dimension involves respect for human dignity regardless of the particular characteristics that a human existence can take, i.e., ethnicity, age, socio-economic level, place of residence, etc. The collective dimension refers to the kind of society that artificial intelligence contributes to creating, i.e., more democratic or more authoritarian societies, economically more just or unjust societies, environmentally friendly societies or societies indifferent to the ecology of the planet, etc.



The promotion and protection of human rights in the new artificial niches

Individual liberties. The risks of the AI's ability to track and analyze our digital lives are compounded by the vast amount of data we produce when using the Internet. With the increasing use of Internet of Things devices and the attempts to shift to "smart cities," people will create a data trail for almost every aspect of their lives. This aggregated data reveals details about our lives. The AI system will use this data to process and analyze them for different purposes, namely, micro-targeted advertising, optimization of public transport, government surveillance of citizens, among others. In such a world, not only are there enormous risks to individual freedoms, but there is also the issue in question of whether or not it is possible to protect these data.

The right to work under fair and favorable conditions. If automation revolutionizes the labor market, large numbers of people may lose their jobs; however, they will have to continue to struggle to support themselves and their families. How is it possible to secure a job and a decent standard of living with such volatility in the labor market?

Online harassment by robots can threaten freedom of speech. Bot accounts that disguise themselves as real users and send disproportionate automated responses to accounts that share a certain opinion can curtail freedom of speech.

But the political organization of democratic states can be put in check by the development of AI as well. Let's consider the following factors.

The various platforms emerge as the new public squares. But they are not public, and information does not flow as freely as in public places (Eslami et al., 2016). "Social bubbles" are created there, where only the information that reinforces our own opinion can be found. On the other hand, these platforms do not seem to be fully responsible for the false information that circulates and harms both users and entire nations, nor do they protect users from so-called trolls and malicious anonymous behavior, even though they can, because it is a very costly process in economic terms.

In a context where wealth and power are being concentrated in an increasingly small number of people, the transhumanist dream could completely alter the social and political life that we know today. If the human improvements that new technologies promote become affordable to only a few, our current social relationships would be radically changed. A longer-lived, healthier, and more capable caste of humans than the rest of humanity would emerge. Thus, transhumanism could direct us towards a new world order governed by elites that have chosen and designed themselves. In a society where these huge differences exist, the demise of democracy seems inevitable.

Let us finally consider an example in which we can find a combination of several of the ethical difficulties that we have outlined throughout this work.

Let's consider the case of an unmanned aerial vehicle (drone) equipped with AI systems designed by a state agency, which is in charge of public security programs in a developing country with a democratic system in consolidation. This device has the ability to collect, process and transmit information in real time in open urban spaces without being detected. It is designed to recognize urban geographic settings, identify people through face recognition, and recognize different kinds of human displacement to distinguish between "habitual" human displacement and human displacement that may suggest the commission of a crime or the like.



In order to assess the way in which technology is linked to society, the notion of well-being that is promoted, the way in which human rights appear and the way in which social practices and citizen coexistence are modeled, one must take into account a set of considerations that range from the decisions made during its design to the nature and characteristics of the agents involved in its development and use. What the technology is designed and developed for and who its users are involves value commitments that may or may not be in tension with the experiences and moral feelings of a given social group.

Technologies like the drone in this imaginary example are not value-neutral solutions. In order to identify these evaluative commitments and calibrate them ethically, at least the following 4 questions should be answered:



- Identification of functions: what tasks does the artifact perform?
- Identification of resources: what material and cognitive resources are needed to make this artifact? Identifying the necessary expert resources contributes to mapping the interests and power structures involved in artifact development.
- Identification of agents: which agents are involved in its development and use? For what reasons do the agents involved develop the technological system in question? Who do they associate with for their development?
- Identification of consequences: what are the consequences for human life, both socially and individually, of the introduction of this device in its environment? Do the social practices that this technology introduces collide with human rights? With which right or rights?

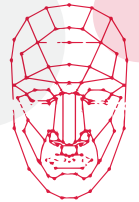
The solution to these questions makes explicit the evaluative point of view on the world that underlies technologies, as well as the ethical tensions that could be involved in devices. Technology is not merely instrumental or the best solution to a practical problem; it is neither innocuous nor morally empty. Let's see how these questions are solved in the case of the imagined example.

The drone - designed by a state agency in charge of public security - has particular functions: recognizing urban geographic settings, identifying people through face recognition systems, and identifying and recognizing different patterns of human displacement. The purpose behind these functions is to facilitate and precipitate decision making by the city police to control mass demonstrations or marches made by citizens.

It is an artifact that fuses control, communication, computer science, identification and recognition. This allows the artifact to locate leadership in mobilized human groups, calculate

the strength of the demonstration, identify displacements and label human behaviors (for example, to identify violent behaviors) in order to develop control strategies in a broad sense, that is, of containment and intervention.

Identifying the necessary material and cognitive resources raises its own problems. For example, which databases will feed the face recognition algorithms? The citizens' identity registration data? The records of citizens with a police record? What behavioral patterns are supplied to the algorithm to recognize dangerous human displacement? What does "dangerous" mean when applied to human movements?



The identification of the agents involved in the development and use of the device entails the explicitation of a set of interests and particular values. In this case, the device is designed by the agency in charge of public security, which implies that a point of view related to citizen control is developed by means of the monitoring of popular concentrations to define their contours and displacement in real time through the surveillance, recording and processing of what happens with the protest. Since it is a state agency in charge of public security, the information collected could be used to open possible files for action to those who participate in the mobilizations - not only their most visible leaders - and transform them into judicial evidence if the state agency considers it convenient.

The point of view transmitted by the drones of our example, equipped with AI and developed by the state agency for public safety, is in tension with the promotion and respect for human rights. Once these artifacts have been introduced for these purposes into social life, is the right to freedom of speech and movement guaranteed, and also to assemble and express one's beliefs without fear of any kind?

The example suggests that the introduction of artifacts such as these in a city environment could change the way in which citizens conceive of themselves and perceive their role in the public sphere, where issues regarding the common good of a human group and their respective practical identities are formulated.

When developers and state authorities justify the technological advantages of drones in terms of security and social control to public opinion, they are positioned as instruments of state power - due to their ability to maintain social order - and leave aside the way in which the exercise of citizen rights is diminished (the ability to express political and social demands on the street).

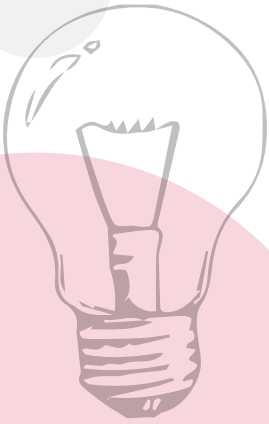
The practices that enable these technologies (monitoring and control of social mobilizations) come into tension with basic ethical contents, which are based on the autonomy of human life and enshrined in the Universal Declaration of Human Rights.

Military drones can be converted into automatic weapons without human control. Let's imagine they are equipped with software that distinguishes objects from human people and then proceed to shoot. But what would happen if instead of a group of terrorist soldiers, the drones are facing three young people hunting rabbits? Should automatic drones be banned - without human control - as chemical weapons are currently banned? Ethical reflection can inform debate in international organizations.



Who is responsible for this technology? Who should be held accountable for their introduction into social life? The State is the main agent in this case and it should be held accountable since it has designed and introduced this technology for a specific purpose: social control. In the case of our example, this has a special impact since it is a society where democratic life is in a consolidation process. This cultural component makes the State look suspicious due to previous historical practices linked to the interruption of democratic political processes. This practical social identity, associated with the fight for human rights, conditions the perception of this technological system by the civilian population, as well as the reasons that drove its creation and the decisions made during its development so that it had some functions but not others. Hence, any assessment is strongly associated with a context with a story.

The decisions that lead to the production of a technological system are permeated with values that act as filters, catalysts and choice motivators in one direction or another. In the hypothetical case at hand, the authoritarian tradition of political life guides the production of a technological system related to special monitoring of the way in which political movements behave on the streets of their cities. In short, the technological system introduced into reality is not only permeated with values, but does not live (or function) in any case in a techno-socio-cultural-political vacuum. A complete understanding of these contexts constitutes a basic condition for understanding what the practical identities at stake are and their respective historical narratives, which will condition and shape a given technology. Consequently, technology should not be analyzed in isolation but in its particular social niches, where it will have to develop and live.



CONCLUSION

Throughout this work, we have pointed out the complexities inherent in the ethical assessment of system developments involving AI. These complexities are tied to fundamental issues for life and human societies that philosophy has sought to unravel from its origins. What counts as a good life? Which actions are morally correct and which are not? What responsibility do each of us have, from our particular place in the public space, in relation to the social and cultural changes in which we are immersed? Who should be held responsible for the problems and conflictive situations we are experiencing? What type of political organization should be promoted and what actions attempt against this way of life? What counts as a just society? What are the limits of private property? What human rights will we recognize as basic and inalienable?

As we have argued throughout this work, all these problems are present in the development and use of AI systems. Questions arise and become complex with each step of the technological innovations that these systems introduce into human life. We have explained the different problems and questions that arise at each step, and we have tried to demonstrate that the responsibility for AI developments to be beneficial for humanity is distributed among all the agents involved: designers, developers, entrepreneurs, government entities and the general population. Shared information, critical reflection and education are essential for all these agents to exercise their rights and obligations in the development of an ethical AI.

Bibliographic references

Aristotle, *Ética Nicomaquea* [Nicomachean Ethics] (5th ed., translation by Antonio Gómez Robledo), Porrúa, Mexico, 1973.

Aristotle, *Obra biológica* [Biological works] (De Partibus Animalium, De Motu Animalium, De Incessu Animalium). Edition under the care of Bartolomé, S. and Marcos, A., Luarna, Madrid, 2010.

Broncano, F., *Mundos artificiales. Filosofía del cambio técnico*, [Artificial worlds. Philosophy of technical change] Paidós, Mexico, 2000.

Burrell, J., How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*. Available at: <https://doi.org/10.1177/2053951715622512>, 2016.

Clark, A. (ed.), *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press, Oxford, 2008.

Clark, A., *Being There: Putting Brain, Body and World Together Again*, MIT Press, Cambridge, 1997.

Clark, A. and Chalmers, D. J., The extended mind, on *Analysis* 58 (1), 1998, pp. 7-19.

Domingos, P., A few useful things to know about machine learning, on *Communications of the ACM*, v.55 n.10, October 2012. [doi>10.1145/2347736.2347755]

Eslami, E., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., Kirlik, A. First I "like" it, then I hide it: Folk Theories of Social Feeds, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, San Jose, California. [doi>10.1145/2858036.2858494]

Frankfurt, H. G., *The Reasons of Love*, Princeton University Press, Princeton, NJ, 2004.
Friedman, B. and Nissenbaum, H. Bias in computer systems, *ACM Transactions on Information Systems (TOIS)*, v.14 n.3, p.330-347, July 1996. [doi>10.1145/230538.230561]

Glenn, T. and Monteith, S., Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections, on *Current psychiatry reports*, 16, 494, 10.1007/s11920-014-0494-4, 2014.

Hill, R. K., What an algorithm is, on *Philosophy & Technology*, 29 (1), 2015, pp. 35-59.

Johnson, J. A., *Technology and pragmatism: From value neutrality to value criticality*, SSRN Scholarly Paper, Social Science Research Network, Rochester, NY, 2006. Available at: <http://papers.ssrn.com/abstract=2154654>

- Kitchin, R., The ethics of smart cities and urban science, on 374 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016. <http://doi.org/10.1098/rsta.2016.0115>
- Markowetz, A., Blaszkiewicz, K., Montag, C., Switala, C. and Schlaepfer, T., Psycho-Informatics: Big Data shaping modern psychometrics, on Medical hypotheses, 82. 10.1016/j.mehy.2013.11.030, 2014.
- Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175-183
- Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175-183
- Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175-183
- Matthias, A., The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6 (3), 2004, pp. 175-183.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L., The ethics of algorithms: Mapping the debate, on *Big Data & Society*, 2016. <https://doi.org/10.1177/2053951716679679>
- Newell, S. and Marabelli, M., Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification', on *The Journal of Strategic Information Systems*, 24, 2015, pp. 3-14.
- Ortega y Gasset, J., *Meditación de la técnica y otros ensayos sobre ciencia y filosofía*, [Meditation on technique and other essays on science and philosophy] Alianza Editorial, Madrid, 1998.
- Pedace, K., *Mente y lenguaje. La filosofía de Donald Davidson, modelo para armar*, [Mind and language. Donald Davidson's philosophy.] SADAF, Buenos Aires, 2017.
- Rawls, J., *A Theory of Justice*, Harvard University Press, Cambridge, MA, 1971 (revised edition, 1999).
- Raz, J., *The Authority of Law. Essays on Law and Morality*, Clarendon Press, Oxford, 1979.
- Romei, A. and Ruggieri, S., A multidisciplinary survey on discrimination analysis, on *Knowledge Eng. Review*, 2014, 29, pp. 582-638.
- Sadin, E., *La silicolonización del mundo*, [The silicolonization of the world.] Caja negra, Buenos Aires, 2018.
- Schermer, B., The limits of privacy in automated profiling and data mining, on *Computer Law and Security Review*, 27, 2011, pp. 45-52. 10.1016/j.clsr.2010.11.009.
- Shalev-Shwartz, S. and Ben-David, S., *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, 2014.

Tene, O. and Polonetsky, J., Judged by the Tin Man: Individual Rights in the Age of Big Data, on Journal of Telecommunications and High Technology Law, August 2013.

Turilli, M. and Floridi, L. The ethics of information transparency, on Ethics Inf Technol, 2009, 11: 105. <https://doi.org/10.1007/s10676-009-9187-9>

Turing, A. M., Computing Machinery and Intelligence, on Mind, 1950, 49, pp. 433-460.

Turkle, S., Reclaiming conversation: The power of talk in a digital age, Penguin Press, New York, 2015.

Tutt, A., An FDA for Algorithms, on 69 Admin. L. Rev. 83, 2017. Available in SSRN: <https://ssrn.com/abstract=2747994> or <http://dx.doi.org/10.2139/ssrn.2747994>

Ward, A. Duke, K. Gneezy, A. and Bos, M. A., Brain Drain: The Mere Presence of One's Own Smartphone Reduces Available Cognitive Capacity, on Journal of the Association for Consumer Research 2, n° 2, April 2017, pp. 140-154. <https://doi.org/10.1086/691462>

Yuste, R., Goering, S., Arcas, B., Bi, G., Carmena, J., Carter, A., Fins, J., Friesen, R, Gallant, J., Huggins, J., Illes, J., Kellmeyer, R, Klein, E., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Ramos, K., Rommelfanger, K. and Wolpaw, J., Four ethical priorities for neurotechnologies and AI, on Nature, 551,2017, pp. 159-163. [10.1038/551159a](https://doi.org/10.1038/551159a).

Zoller, Y., The costs of overprotecting the young - iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy -and completely unprepared for adulthood- and what that means for the rest of us by Jean M. Twenge, on The American Journal of Psychology, 2019,132, pp. 115-119.

Disclaimer. The opinions expressed in the publication are the sole responsibility of the authors. Said opinions do not intend to reflect the opinions or perspectives of CETyS or any other organization involved in the project.