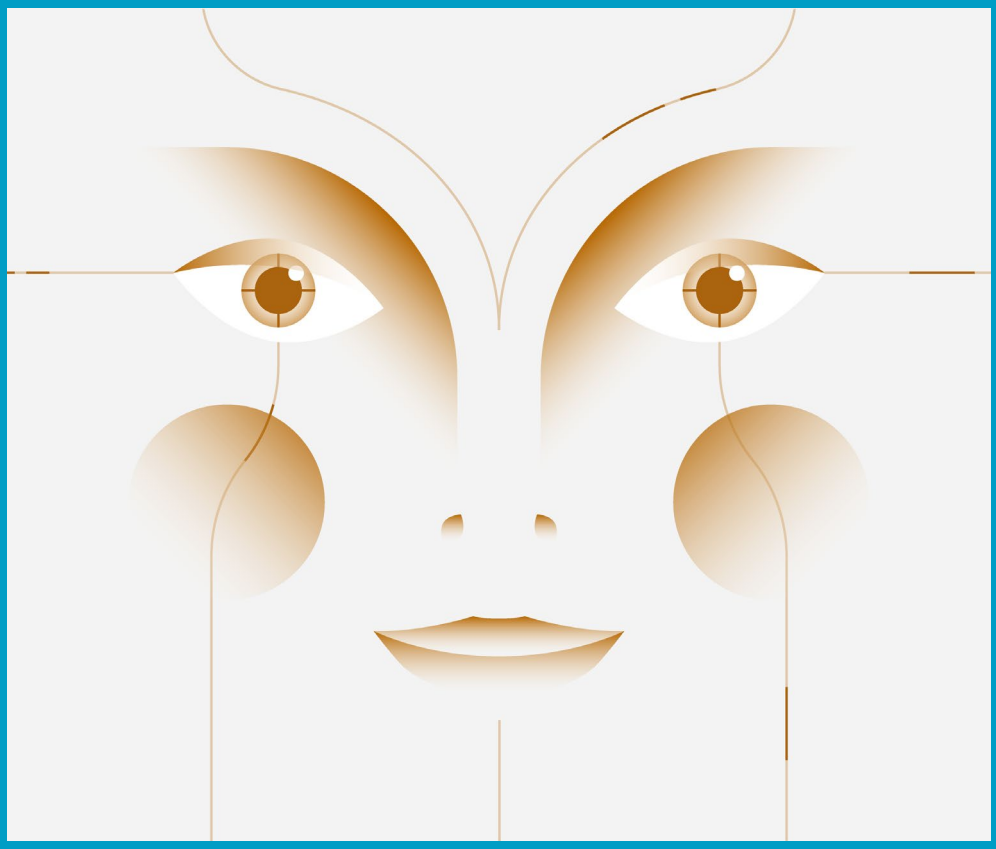
Building TrusTee

The World's Most Trusted Robot

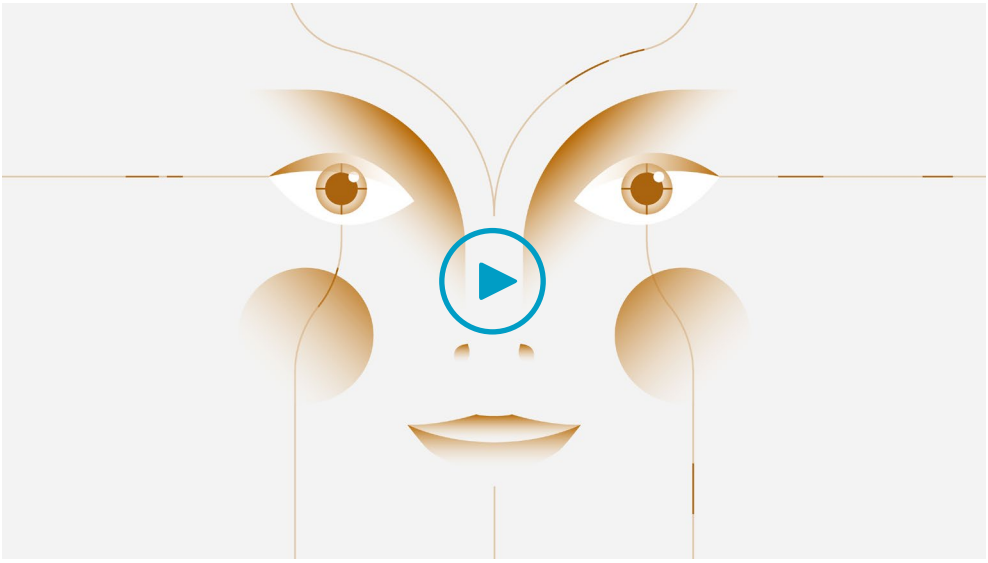
Ton Wilthagen
Marieke Schoots



Tilburg University

Building TrusTee

The World's Most Trusted Robot



Ton Wilthagen[1]
Marieke Schoots

Tilburg University Impact Program
November 2019

---

1 We thank all colleagues from Tilburg University and various other partners that have provided input, thoughts, suggestions and questions during the writing and production of this essay. The essay was prepared for the occasion of the "Man-Machine & Values" seminar organized by the Tilburg University Impact Program on November 14, 2019. User note: this essay contains many hyperlinks that lead you to further information and references, including a short animation of TrusTee. Just click on the links (or play button). Moreover, if you would like to enjoy some suitable music while reading, go to the TrusTee playlist on Spotify where you find handpicked tracks that either musically or text-wise relate to the topic at hand. TrusTee can be contacted via Twitter and Instagram: @robottrustee

TrusTee

The world's most trusted robot.

"The computer can learn on the basis of its experiences and this means that it can take into account the outcomes of earlier decisions made under similar conditions, when making a decision (...) A computer, a product of the human mind, will never be able to take over human thinking. The computer can nevertheless be of the utmost significance for our thinking. After all, the computer forces us to a higher level of accuracy in formal thinking and in articulation".[2]

2 Max Euwe, mathematician and first and only Dutch world champion chess player, in his inaugural lecture 'Can computers think?' (Kunnen computers denken?) as professor in the Methodology of Automatic Data Processing, at the predecessor of Tilburg University, on October 29, 1964.

Contents

1. There's a new kid in town

Robotization, digitalization, and artificial intelligence (AI) are developing at a fast pace and penetrating and influencing all aspects of life and society. Robots are leaving the cage in which industrial robots have been functioning for quite some time already. They are becoming "cobots," i.e., collaborative robots. Robots do not always look like us humans. Besides humanoid robots, various forms and appearances of the new technology exist, including softbots, chatbots, drones, and an unlimited number of applications of algorithms and artificial intelligence. It is going to be the first time in history that humans will intensively interact and live together with non-human actors, a.i. artefacts that can operate autonomously, are very powerful and intelligent, and can actually learn and improve themselves, either through machine learning or deep learning.

So far, we as humans have been living with and relating to other humans from similar or different cultural and ethnic backgrounds, part of whom are family or friends and (domestic) animals. In addition, a part of the human population relates to Gods or other Supreme Beings. We have always had some difficulties in understanding "the other" who, besides our children, we did not create, let alone manufactured.

Now a kind of "superintelligence"[3] is emerging, and various authors argue that the point of singularity, where robots and AI outsmart us as humans, may not be so far away[4]. As Elon Musk stated: "Robots can do everything better than we can."[5]

3 Bostrom, N. (2017).
4 It is interesting that many people that are earning or have earned a lot of money with intelligent technology also strongly warn against the effects of this very technology. Or even go as far – after retirement – to make a plea for a robot tax in the case of Bill Gates (Delaney, K.J., 2017)
5 Clifford, C. (2017)

Box 1

Professor and world champion chess player Max Euwe versus the computer

The Dutchman Max Euwe (1901-1981) was a brilliant mathematician and the first and only Dutch world champion chess player. He was also a professor of automation/computer science at the universities of Tilburg and Rotterdam. Being aware of the fast increasing importance of computers in society, he wrote books for teachers in vocational education and contributed to a national TV course (Teleac) titled *Mastering the computer (Hoe word ik de computer de baas?)*, as early as 1974.

For Euwe, chess represented the ways in which humans (or at least chess grandmasters) differ, and will continue to differ, from computers. He considered hunches, inspiration, and intuition typical for how we as humans think and operate: something that computers would never be able to imitate. Thus, Euwe was convinced that the computer would never be able to beat an outstanding chess player, as opposed to checkers, which he considered a different game.

At Euwe's retirement in Tilburg in 1971, he reluctantly admitted that it was getting more and more difficult for him to beat a chess computer. He did not live long enough, maybe for the better, to see that, 25 years later, on February 10, 1996, IBM's Deep Blue computer beat world champion chess player Garry Kasparov.[6]

A number of years later, in 2017, Tilburg University professor Jaap van den Heerik, also a chess player and also in a farewell lecture, concluded that "intuition can be programmed."[7]

The developments we are now sparking and, at the same time, observing and experiencing may be considered a blessing in disguise or more precisely a blessing in device. Anyway, the "robot revolution" will have impact.[8] As Kranzberg's first law states: "Technology is neither good nor bad; nor is it neutral."[9] Robots and AI can contribute to the "social good," to a "good society" by making life safer, easier, better, healthier, and more efficient and productive.[10] New jobs will appear that people can take up.

6 In line with Minsky, M. (1961, p. 9): "(...) systems like chess, or nontrivial parts of mathematics, are too complicated for complete analysis. Without complete analysis, there must always remain some core of search, or "trial and error." So we need to find techniques through which the results of incomplete analysis can be used to make the search more efficient."
7 Van den Herik, J. (2016)
8 Hudson, J. (2019)
9 Sacasas, L.M. (2011)
10 Omohundro, S. (2014)

Furthermore, people's direct participation in decision-making and politics can be enhanced, so we can all have a real-time say in society and shape our future. Additionally, smart technology may have benign effects on our living environment and the condition of our planet.

At the same time, however, this technology may turn out to be a devil in many disguises, threatening privacy and other human rights and contributing to new "techno" forms of disciplining, control, and even repression. It may drive people out of jobs (which is already happening) and, in general, lead to alienation and dehumanization. We might become (too) strongly dependent on robots and AI.

Box 2 explains how robotization works out in the employment domain.

Box 2

Job destruction or job creation?

"Robots taking our jobs," "Humans Need Not Apply". The impact of robotics, AI, and automation in general is one the most widely discussed and feared issues.[11] Without jobs, people will not have much income security and lose the most important way of participating in society. If you meet someone at a party you do not know yet, the first question is often "What do you do?" Many early estimations of jobs to be lost to robots now seem exaggerated, such as that some 40 percent of all jobs in the US would be lost in the next twenty years. Later estimations speak of 14, 10, or even 5 percent. It is very hard to predict. There is a general consensus that mostly workers with lower-middle levels qualifications in manufacturing and in the financial services are vulnerable to job loss.

This is the so-called "skills-biased technological change" notion. More recently, another notion has been added, that of "routine-biased technological change," meaning that the risk that a robot takes your job depends on the amount of time you spend on routine tasks within your job.[12] This implies that also people that perform higher-level cognitive jobs with many routines (like accountants and lawyers) are at risk. Alternatively, people who do manual, non-routine jobs (like a waiter in a bar) would have less to worry about.

11 Ford, M. (2016). See also Inglehart, R.F. (2019)
12 Chui, M. et al. (2016)

Therefore, we cannot properly calculate the net balance. Nevertheless, what we observe is that labor market participation rates in many countries have been rising over the past decades, in spite of the internet booming and the growing numbers of industrial robots.

The original Czech meaning of the word "robot" equals "slave", but, ironically, in a worst-case scenario, humankind might end up being enslaved by robots. This is the dystopian scenario of Aldous Huxley's famous 1932 novel Brave New World. In that case, robots and AI might turn out to be our "final invention".[13] At a very early stage, science fiction writer Asimov suggested three laws on robotics. See box 3.

Box 3

Isaac Asimov's Three Laws of Robotics[14]

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

The obvious thought of building a so-called red button into the design of complex technology, to stop things when they go out of control according to human operators, is probably illusory. Within highly complex technological systems, abruptly pushing the button may lead to chaos and the loss of functions that we would like to see continued.[15]

All these possible scenarios, fears, and hopes have been highly anticipated by the arts, firing our imagination: through science fiction movies, books, paintings, and sculptures. Art has inspired science, as science has inspired art in this domain. In this essay, we will sometimes refer to these artistic expressions.[16]

13 Barrat, J. (2013)
14 Gunnoo, H.A. (2019)
15 Arnold, T. & Scheutz, M. (2018)
16 Tilburg University was a partner in Robot Love, an interactive experience exploring love between humans and robots with a wide array of expositions and activities. A large number of artworks were made specifically for Robot Love. The exposition took place from September 15 until December 2, 2018 at the Campina Milk Factory in Eindhoven, the Netherlands. More than 60 artists, designers, and scientists asked themselves whether robots are capable of love and, vice versa, whether we can love them too.

The relationship between technology and humans is complex and the development of artificial intelligence brings new challenges, requiring a debate about the social and ethical implications.[17]

As Halman and Sieben write: "Values gained momentum again recently, triggered by the discussion about robotization, digitalization of society, and the growing role of artificial intelligence."[18] The critical issue and solution with regard to human-technology interaction is called "value alignment". As Vamplew and others put it, "As the capabilities of artificial intelligence (AI) systems improve, it becomes important to constrain their actions to ensure their behaviour remains beneficial to humanity."[19] The assumption is that if we are able to design and program robots and AI in such a way that they align in every operation and effect with what we consider human values—in manners, process and outcome—we need not be too concerned, and we will experience great benefits.

Therefore, the central question in this essay is: *how do we build the world's most trusted robot, what is needed in doing so, and, in particular, what is the role and relevance of the social sciences and humanities in view of the significance of values in this design development process?* For this purpose, we will refer to insights and rely on sources from various disciplines.

17 Halman, L. & Sieben, I. (2019)
18 Halman, L. & Sieben, I. (2019, p. 5)
19 Vamplew, P. et al. (2017, p. 27)

2. Introducing TrusTee

Let us call this most trusted and social robot "TrusTee," which stands not just for robots but also for any manifestation of intelligent, autonomous, and self-learning technology. It could also stand for Tilburg University Social and Trusted Robot. At Tilburg University, the Netherlands, we are highly committed to building TrusTee by being part of a worldwide effort and partnership with other research centers, political bodies, NGOs, and companies. Just to make sure, currently, TrusTee is an image, an imaginary friend a vision, an approach but certainly a prospect–not one single super robot or AI application already under construction.[20]

Evidently, we also chose the name TrusTee because we need future and intelligent technology to be our trustee. The Merriam-Webster online dictionary defines a trustee as "a natural or legal person to whom property is legally committed to be administered for the benefit of a beneficiary (such as a person or a charitable organization)." We want to be able to entrust TrusTee with the future of our societies and planet.

Much research is already going on regarding the ethical aspects of robotics and the role of trust in human-robot interaction. An early example is the 1991 paper by James Gips, Towards the Ethical Robot, *Can we trust robots* by Mark Coeckelbergh from 2011. and *Can You Trust Your Robot?* by Hancock et al., from the same year.[21]

20 Right after the completion of the first draft of this essay, a book by was published, The Fall of the Human Empire: Memoirs of a Robot, authored by Charles-Edouard Bouée (expected, 2020). This book tracks down the history of artificial intelligence from the point of view of a robot called Lucie. Lucie narrates her adventures and discloses the mysteries of her journey with humans, reflecting on what developments in A.I. may mean for both humans and robots.

21 Gips, J. (1991), Coeckelbergh, M. (2011), Hancock, P.A. et al. (2011)

We also refer to Kranzberg's Fourth Law[22]: Technology might be a prime element in many public issues, but nontechnical factors dominate in technology-policy decisions and implementation, and various complex sociocultural factors, especially human elements, are at play, even in what might appear purely technical decisions. Within the TrusTee project, we mobilize and involve the social sciences and humanities (SSH), economics, law, public administration and governance, communication and cognitive sciences, data science, psychology, sociology, philosophy, and theology.

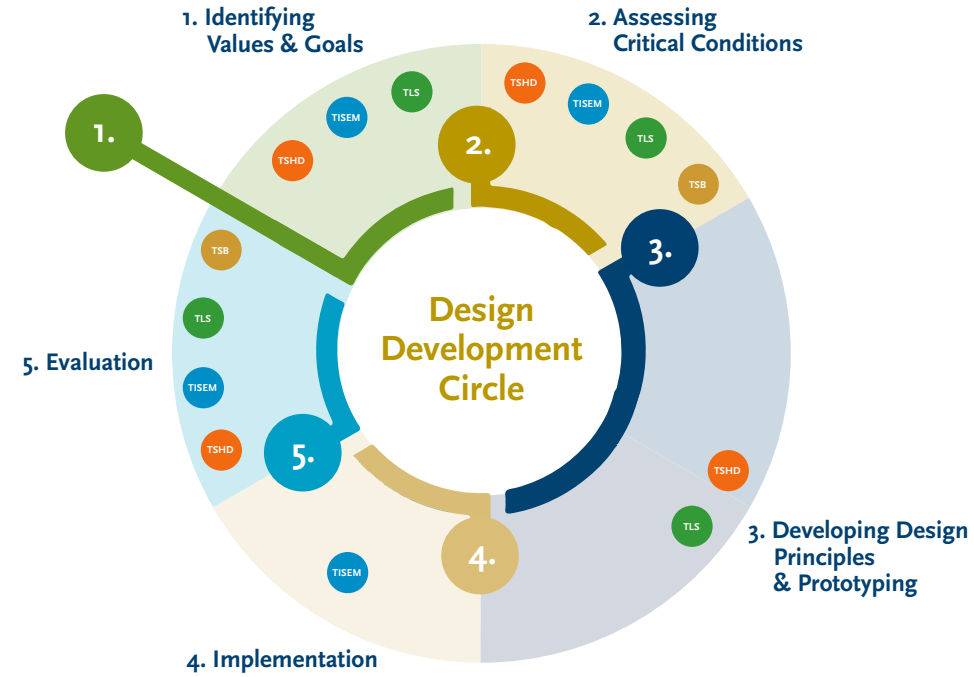
Technological and social innovation need to go hand in hand, an inference drawn by a growing number of authors and organizations.[23] In the Netherlands, as elsewhere, the SSH are emphasizing and claiming that they have a right to play in society's major challenges and missions as well.[24] In the traditional view on the SSH in relation to technology, these disciplines are believed to have had a limited, secondary role with a large emphasis on issues like facilitating the acceptance and user-friendliness of technology or more legal issues like privacy. Seen from a more contemporary viewpoint, the SSH take a more fundamental and primary position, seeing technology development as a means to an end rather than as an end in itself.

Naturally, the SSH will work closely together, as we do already, with STEM -partners (Science, Technology, Engineering, and Mathematics), including other universities, research and technology institutes, and tech companies in and outside the Netherlands. A good example is the co-founding of the Jheronimus Academy of Data Science in the City of Den Bosch by Tilburg University and the Eindhoven University of Technology.

The TrusTee endeavor aptly fits the ambitions and aims of the Tilburg University Impact Program, where we work to advance society from the perspective of "Science with a Soul."[25] The project is neatly positioned amongst our three impact themes, derived from the Dutch National Science Agenda (*Nationale Wetenschapsagenda*): Empowering the Resilient Society, Enhancing Health and Wellbeing, and Data science for the Social Good. It also builds on the specific scientific strengths and expertise of our university (see further below) and our role in the Digital Society program of the Association of Universities in the Netherlands *(VSNU)*. The latter program aspires to making the Digital Society even *more* humane than the analogue society–so not less.

22 Sacasas, L.M. (2011)
23 Dolling, L. (2013), Prinster, R., (2017), Glaser, L.B. (2016)
24 Read more about sector cooperation and the implementation plan for the Netherlands (both in Dutch).
25 Wilthagen, T. et al. (2017)

We can distinguish 5 steps in building TrusTee, as portrayed in the design development circle below:



TSHD Tilburg School of Humanities and Digital Sciences
TISEM Tilburg School of Economics and Management
TSB Tilburg School of Social and Behavioral Sciences
TLS Tilburg Law School

For further information on the scientific strengths please click on the buttons in the picture.

For each step, we identify the scientific strengths or competences that are required and their availability at Tilburg University.[26] In this essay, we focus on the role of values in building trustful and responsible robots and AI. As John Havens has put it, we need to focus on "Heartificial Intelligence"[27] or "responsible AI" as Victoria Dignum words this goal.[28]

26 Scientific Strengths, Tilburg University, (2018).
27 Havens, J.C. (2016)
28 Dignum, V. (Expected December 2019)

1. Identifying values and goals key to the development of TrusTee
We start out by collecting all the major treaties, charters of rights, conventions, basic laws, the United Nations Sustainable Development Goals, et cetera. In theory, all sorts of ethical codes and moral systems could be uploaded into TrusTee, but we firmly believe that values that have been codified in democratically legitimated legal rules and agreements are the values that have the most gravitas, universality, and timelessness and are the most thoroughly justified.

2. Assessing the critical conditions
At the same time, we will have to perform a meta-analysis of the major empirical findings that shed light on the issues where we, as humans, have or have not succeeded to implement these values effectively. What has worked, what has not—what have been and still are the critical conditions for effecting core human values and goals?

3. Developing the key design principles and prototyping
We research, identify, and define—from an interdisciplinary perspective—the key design principles that represent and constitute the most important human values and their critical conditions and implications when it comes to robots' and AI's decision-making, routines, actions, operations, and effects.[29] This includes very general human-robot interaction design principles, such a suggested in Asimov's robot laws. The design principles should be operationalized in such a way that they can actually guide the decision-making and actions of robots and AI based on "values by design." Design principles, coding, and algorithms need to be open and 100% transparent from the very start.

4. Implementation
Evidently, TrusTee needs to be aware of and detect relevant situations in which to deal and work with (rather than merely apply) the key design principles and codes through sensory observation. AI and robots will most likely be able to move beyond the typical human scope of observation as much more data and information can be collected, processed, and combined.

5. Monitoring and learning
It goes without saying that TrusTee's actions/non-actions, and the impact thereof, need to be monitored carefully across all domains of life and society. This, again, is also a major task for the Social Sciences and Humanities in collaboration with STEM. Actions and subsequent monitoring should lead to learning. Here, the good news is that TrusTee will also be able to learn by itself through machine learning and deep learning. However, learning will also remain a co-creation/joint effort between humankind and TrusTee.

29 On interdisciplinary approaches and team science see Wilthagen, T. (2018)

3. Moral agency or not?

Values come into play once we link technology to morality. If the technology does not have any moral significance, we do not have to discuss values and morality, at least not with robots and AI—we should just have this discussion among ourselves, as humans. Speculation about robot morality is nearly as old as the concept of the robot itself.[30] In the current literature and public debate, there is quite some confusion on moral agency and technology.

There is also an 'amoral', or rather, practical position in the debate among some researchers and, in particular, among officials of ministries of Economic Affairs in European countries. The argument goes as follows: Let us just develop the technology first, to make sure we are not outcompeted by China and the US who have much less scruples than we do. If we bother too much now, we will slow down innovation and, anyway, we cannot yet foresee all the possible applications of what we are developing.

A first position on morality seems to be that robots and AI do not have any moral impact by themselves, i.e., they are just tools that, like in the old days, humans apply. As with a hammer, you can either build a house with it, or beat someone to death. This—moral—decision is up to the user, i.e., the human actor. This is more or less the position Etzioni and Etzioni assume. Consequently, they find that "a significant part of the challenge posed by AI-equipped machines can be addressed by the kind of ethical choices made by human beings for millennia. Ergo, there is little need to teach machines ethics even if this could be done in the first place."[31] "Ethic bots" are seen as possible mediators between, e.g., a car-owner and a self-driving car, as they will translate the owners' wishes to the car.[32]

30 Leenes, R. & Lucivero, F. (2014)
31 Etzioni, A. & Etzioni, O. (2017, p. 403)
32 Etzioni, A. & Etzioni, O. (2017, p. 415)

A second point of view is that the operations and workings of the technology do have moral implications, with humans programming these implications—analysis, decisions, and acts. For example, currently AI is already contributing to the selection of CVs of candidates in recruitment procedures. If we design an algorithm that throws out every candidate over 55 years of age because we find that this group lacks productivity, we take this moral decision and incorporate it in the algorithm. The consequence here is that the recruitment officer, if still in place, will never look at a CV from a 55+ applicant because this has been predetermined. This is why there is an enormous debate on transparency in developing AI and robots.[33] New technologies intersect with old prejudices.

Box 4

The Amsterdam Tada City manifest on transparent algorithms

Recently, Amsterdam, as a city, has decided it requires transparency of every algorithm that is at work in the city, e.g., in the world of Airbnb. In the so-called Tada City manifest, it is formulated as follows:

Data: a promise for life in the city. Data enables us to tackle major problems modern cities face, such as making them cleaner, safer, healthier... but only as long as people stay in control of the data, and not the other way round. We—companies, government, communities, and citizens—see this as a team effort and want to be a leading example for all other digital cities across the globe.

The Amsterdam initiative is welcomed by, among others, Cathy O'Neil, who authored a book with the telling title *Weapons of Math Destruction*.[34] It can be considered an interesting example of "keeping society in the loop" by programming a societal contract by means of artificial intelligence.[35]

33 A current case in point is the face-scanning algorithm designed by the recruiting-technology firm HireVue. According to Drew Harwell, Washington Post, this system uses candidates' computer or cellphone cameras to analyze their facial movements, word choice and speaking voice before ranking them against other applicants based on an automatically generated "employability" score. It increasingly decides whether people deserve the job. Some AI researchers argue that the application is not rooted in scientific fact. Analyzing a human being like this, they object, could result in penalizing nonnative speakers, visibly nervous interviewees or anyone else who doesn't fit the model for look and speech. Another example concerns the System Risk Indication (SyRi) that has been developed as part of the Dutch Anti-Fraud System and is now under attack. Dutch governmental institutions are allowed to cooperate in intervention teams to detect tax and allowance fraud and noncompliance with regulations in the field of employment and social security. SyRi can compare risk profiles with real person cases and indicate the relevance for further investigation. UN-rapporteur for Human Rights Philip Alston has expressed his concerns about this system in a letter to a court in The Hague. He concludes that the system violates human rights because it discriminates against people with limited financial means and a migration background.

34 O'Neil, C. (2016)

35 Rahwan, I. (2018)

In various publications, there is talk of AMAs, Artificial Moral Agents. Thomas Cheney writes in a blog:[36]

An 'AMA', however, is more than just a programme executing commands, it takes actual decisions, it makes moral choices, even if it is not 'conscious' or 'sentient' (...) an AMA goes beyond simply being an 'autonomous intelligent' system to one that makes moral decisions. AMA refers to systems that are more than just excellent computers, but systems that actually 'think', that should therefore be responsible for their decision.

Therefore, a third position states that robots and AI are or at least will become full moral agents, in the sense that they will develop, learn, and ultimately acquire an independent and superior moral status. This is again the superintelligence hypothesis.

A fourth take on the issue is to actively attribute "artificial morality" to autonomous technology (along the line of "artificial consciousness") and study how this morality could be designed and promoted.[37]

Should the technology be seen as a moral actor and, therefore, be protected and attributed rights, e.g., the right not to be mistreated or destroyed, as laid down in Asimov's third law? The question then is how to assess the degree of morality. These questions are rather urgent, e.g., when it comes to the liability of (semi-)autonomous systems, including self-driving cars that cause an accident.

One of the solutions is offered by ethical behaviorism that states that morality is simply a matter of behavior. For example, Danaher holds that robots can be attributed moral status if they behave more or less "performatively equivalent" to humans who have important moral status.[38] Intentions and motives are not included in the equation; there is no need to assess moral reflection based on Kantian ''autonomy'' or ''will.''[39] The argument is that the performative threshold that robots need to cross in order to be afforded moral status may be fairly low and that they could soon be welcomed in the moral circle. In this case, the unthinkable happens and Asimov's third law comes into force: robots become legal persons and should be attributed "robot rights."[40]

Currently, in law, it is assumed that human and corporate actors make decisions, not technology. Teubner[41] starts with the description of a case from 1522 in Arlun, where rats were put to trial and concludes that there is "no compelling reason to restrict the attribution of action exclusively to humans and to social systems (...). Personifying other

36 Cheney, T. (2017)

37 See also Allen, C. et al. (2005)

38 Lecture of John Danaher based on Danaher, J. (2019)

39 Wallach, W. et al. (2010, p. 456)

40 Gunkel, D.J. (2018)

41 Teubner, G. (2006, p. 502)

non-humans is a social reality today and a political necessity for the future."

These questions are rather urgent, e.g. when it comes to the liability of (semi) autonomous systems, including self-driving cars that cause an accident.[42]

In an Open letter to the European Commission Artificial Intelligence and robots the signatories protest against the creation of a legal status of an "electronic person for autonomous", unpredictable and self-learning robots", as the idea rests on "incorrect affirmation that damage liability would be impossible to prove" and on an overestimation of the current capabilities of robots.

It is contended that from an ethical and legal perspective, attributing a legal personality to a robot is inappropriate whatever the legal status model:

a. A legal status for a robot can't derive from the Natural Person model, since the robot would then hold human rights, such as the right to dignity, the right to its integrity, the right to remuneration or the right to citizenship, thus directly confronting the Human rights (...)
b. The legal status for a robot cannot derive from the Legal Entity model, since it implies the existence of human persons behind the legal person to represent and direct it. And this is not the case for a robot.
c. The legal status for a robot cannot derive from the Anglo-Saxon Trust model also called Fiducie or Treuhand in Germany. (...) It would still imply the existence of a human being as a last resort – the trustee or fiduciary – responsible for managing the robot granted with a Trust or a Fiducie.

In September 2010, the EPSRC (Engineering and Physical Research Council, UK) suggested the following principles of robotics to complement Asimov's laws (see box 3). The EPSRC felt the need to address the issue because in their opinion Asimov's laws are "inappropriate because they try to insist that robots behave in certain ways, as if they were people, when in real life, it is the humans who design and use the robots who must be the actual subjects of any law"

- Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security;
- Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy;
- Robots are products. They should be designed using processes which assure their safety and security;
- Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent;
- The person with legal responsibility for a robot should be attributed;

42 Tjong Tjin Tai, E. et al. (2018)

Another strongly opposing view is expressed by Van Wynsberghe and Robbins, who wish to shift the burden of proof back to the machine ethicists and demand good reasons from them to build AMAs. Until this is done, the development of commercially available AMAs should not proceed further.[43]

From the latter perspectives, our TrusTee robot would still need the back up of a human being in order to be allowed to act or not act. It is our view that robots and AI already have moral significance, already act and decide in moral ways, even if today this is still predominantly determined by a human agency. Therefore, the issue of value alignment is an urgent one. The biggest consequences of robots and AI is that we as humankind are urged to reflect on our own morality, our own values. As Havens[44] puts it: "But how will machines know what we value if we don't know ourselves."

We cannot just study technology without engaging in self-examination. It is as if all of a sudden aliens arrived on our planet and asked the existential question: Who are you and what do you want? Therefore, the rise of robots and AI is to a great degree a "man in the mirror" situation. Robots and AI ask the question posed in a track by 60s New York rock band The Velvet Underground: "I'll be your mirror, if that's what you want."

The difference with alien intruders is that robots and AI are entering our lives gradually, although quite fast. This means we have a bit of time but probably not much. As said previously, contrary to raising our children, we do not have the length of a childhood to raise and socialize an individual robot, transmitting values, fall and rise, having them acquire qualifications, and then letting go.[45]

Besides, our societal systems of value transfer do not apply (yet) to the new technology. Talcott Parsons theorized this function in his AGIL[46] paradigm, where the I stands for Integration and harmonization to ensure that a society's values and norms are solid and sufficiently convergent. Moreover, the L refers to latency, or "latent pattern maintenance" aimed at warranting the integrative elements of the integration through institutions like family and school, which mediate belief systems and values between an older generation and its successor.[47]

43 Van Wynsberghe, A. & Robbins, S. (2019, p. 719)
44 Havens, J.C. (2016, p. xix)
45 This perspective of educating robot similar to children was already outlined by Alan Turing: Turing, A. (1950)
46 Adaptation, Goal attainment, Integration and harmonization, Latency
47 Parsons, T. (1991)

All human-technology interaction in the robot and AI age boils down to an inverse variant of the “comply or explain” regulatory approach. If we want the technology to comply with our standards, we will have to explain the standards to the technology. Robotics and AI applications will push us to put our cards on the table, to level with the technology. We can no longer leave all things we value implicit; we will have to be much more explicit.

The question here of course is: Do we know and agree what those values are, and are we capable of pursuing “human values by design”?[48]

48 Havens, J.C. (2016, pp. 190-191)) is positive about the feasibility of this and proposes a framework for it.

4. Let's talk values with TrusTee

It seems that we as humans cannot avoid or escape "talking" values with autonomous and intelligent technology. This leads to two major and complex questions: 1) What are values 2) How can we infuse technology with values? We should act now, as we as humans are bound by Amara's law: we tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run. We over-trust things we do not understand.[49] We should observe the precautionary principle, which, according to a European Parliament Think Tank, enables decision-makers to adopt precautionary measures when scientific evidence about an environmental or human health hazard is uncertain and the stakes are high. The current state and performances of robots and AI are not a measure of their potential and impact.

Starting with the first question, we can turn to a long scientific tradition on identifying and researching values from many disciplinary angles. However, first we have to answer the fundamental question whether science has any right at all to say anything about values and morality. This question is the topic of Sam Harris' book *The Moral Landscape. How science can determine values*.[50]

Harris rejects the idea that values are exclusively the domain and jurisdiction of religion. He contests the opinion that science can only tell us how we are and not how we should be. His concept of morality centers on well-being: "Once we see that a concern for well-being (...) is the only intelligible basis for morality and values, we will see that there *must* be a science of morality (...)."[51]

49 Fry, H. (2019)
50 Harris, S. (2012)
51 Harris, S. (2012, p. 44)

Harris also has strong thoughts about the most adequate methodological approach: "a scientific account of human values, - i.e. one that places them squarely within the web of influences that link states of the world and states of the human brain (...)."[52] This, Harris writes in his conclusions, has far-reaching implications:

If our well-being depends upon the interaction between events in our brains and events in the world, and there are better and worse ways to secure it, then some cultures will tend to produce lives that are more worth living than others; some political persuasions will be more enlightened than others; and some world views will be mistaken in ways that cause needless human misery.[53]

What are values?
Values cannot be observed, at least not directly, and neither can they be touched. So how do different disciplines study values? Halman and Sieben provide the following answer.

To the social sciences, values are considered to be crucial factors in everyday life, although they all study values using a distinctive theoretical perspective. In economics the value of products, goods, and services are studied in terms of their utility. The economic theory of value is usually equated with the theory of price. In psychology, values are regarded the motivations for behaviors. In sociology, values are considered to be social standards or criteria that can serve as selection principles to determine a choice between alternative ways of social action. Sociologists are interested in values as far as they are inherent to social systems, i.e. are culturally or structurally determined, and influence the orientation of collectivities.[54]

From sociology, we can also learn an important distinction between values and norms, terms that are often used interchangeably. Values are more abstract and general standards, ends to achieve; something that we find very important. Values are often expressed in one word: freedom, equality, solidarity, safety, et cetera. Norms are specific prescriptions—rules and expectations—for behavior in social situations, certain means to realize a value. A value can be addressed and effected through various norms, not just one. Privacy is a value, keeping a distance from the person standing before you, while waiting in a queue in a post office, is a norm.

Some social norms are legal norms, but you can be called, evaluated, judged, and sanctioned based on all norms, with legal norms possibly leading to legal consequences, e.g., fines or imprisonment, and social norms to informal social consequences, including disapproval or exclusion from a group. In researching human values, we should also pay attention to the fact that many values are also defined as (human) rights and that these rights are good

52 Harris, S. (2012, p. 25)
53 Harris, S. (2012, pp. 243-244)
54 Halman, L. & Sieben, I. (2019, p. 27)

indicators of values. The value of privacy has an equivalent in the right to privacy.

Arguably, we should not just teach robots and AI values, but also inform them on the norm level, unless they are able (and willing) to derive the proper norms to implement the values at stake by themselves.

Within their theory of human development, Welzel, Inglehart, and Klingemann conceptualize the following linkage between individual resources, emancipative values, and freedom:

Socio Economic development gives people the objective means of choice by increasing individual resources; rising emancipative values strengthen people's subjective orientation towards choice; and democratization provides legal guarantees of choice by institutionalizing freedom rights (...) the linkage between individual resources, emancipative values and freedom rights is universal in its presence across nations, regions and cultural zones; that this human development syndrome is shaped by a causal effect of individual resources and emancipative values on freedom rights; and that this effect operates through its impact on elite integrity, as the factor which makes freedom rights effective.[55]

The authors are then able to explain value change because of socioeconomic development when expanding markets and social mobilization diversify and intensify human activities, such as commercial transactions and civic exchanges.

In the psychology of human values, values are linked to psychological needs, feelings, motives, traits, habits, and of course behavior. Moreover, psychology looks at the kind of psychological information values contain and to the psychological resources that can be derived from considering a value important.[56] Values "help to organize our likes and dislikes."[57]

With regard to autonomous and intelligent technology, an important lesson can be derived from what Maio[58] calls the problem of introspection: (1) some people, say certain clients or designers of technology and in our case also machines (e.g. the HAL computer in *2001 A Space Odyssey*), may not be willing to state their true values or (2) are unaware of their true values.

55 Welzel, C. et al. (2003, p. 341)
56 Maio, G.R. (2017, p. 51)
57 Maio, G.R. (2017, p. 127)
58 Maio, G.R. (2017, p. 277)

Finally, Maio[59] is right in drawing our attention to the point that not all moral judgements are obviously related to (personal) values and that moral judgements might not always indicate a threat to certain values. This means that value alignment with regard to human-technology interaction is not by definition the same as moral judgment alignment.

The study of ethics and human values has been the topic of philosophers, moralist, theologians, and later, as explained above, of sociologist, anthropologists, economists, and psychologists. Recently, neurobiology has also carefully moved into this field.

Damasio[60] discusses the traditional answer from cognitive sciences and neurobiology to the question about the origin of the values that enable us to make moral judgments, e.g., about good and bad behavior. This answer entails that a historical process of value construction has taken place, based on the extraordinary development of human intelligence being further perfected and transmitted through generations in view of human interactions and the creative reasoning over these interactions.

According to Damasio, there may already have been "antecedents" for the intelligent construction of human values—a biological blueprint already present in non-human species and early humans. "We simply wish to suggest that the construction was constrained in certain directions by preexisting biological conditions."[61]

Those preexisting biological conditions are then defined as a part of the "life regulation system" or homeostasis that objects to conditions of operation leading to disease and death and looks for conditions that lead to optimal survival. "It's a demonstrable fact that what we usually call good and evil is aligned with categories of actions related to particular ranges of homeostatic regulation (...) What we call good actions are, in general, those actions that lead to health and well-being states in an individual, group or even a species. What we call evil (...) pertains to malaise, disease or death in the individual, the group or the species."[62]

The relevance and consequences of this neurobiological perspective for building trusted and social robots are probably hard to consider at this moment. Do we first need to provide TrusTee with a life regulation system so that it has the neural grounding for value-alignment and will be able to ground the values it is expected to adhere to and help to realize? Or is installing a life regulation system in robots a recipe for trouble, as displayed in many science fiction novels and movies, inviting robots to rebel and protect themselves, even against Asimov's first and second law?

59 Maio, G.R. (2017, p. 285)
60 Damasio, A. et al. (2005)
61 Damasio, A. et al. (2005, pp.47-48)
62 Damasio, A. et al. (2005, p. 48)

5. How can we communicate values with TrusTee?

A further crucial question is how we can communicate values with intelligent technology. As Wallach and Allen[63] rightly observe, values are not a soft thing and we really need to have this conversation:

Some engineers may be tempted to ignore or dismiss questions about values as too soft, but this will not make them go away. Systems and devices will embody values whether or not humans intend or want them to. To ignore values in technology is to risk surrendering their determination to chance or some other force.

Box 5

The 2016 Microsoft twitter experiment[64]

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay — a Twitter bot that the company described as an experiment in "conversational understanding." The more you chat with Tay, said Microsoft, the smarter it gets, learning to engage people through "casual and playful conversation."

63 Wallach, W. & Allen, C. (2010, p. 39)
64 Source and cited from The Verge, Vincent, J. (2016)

Unfortunately, the conversations didn't stay playful for long. Pretty soon after Tay launched, people starting tweeting the bot with all sorts of misogynistic, racist, and Donald Trumpist remarks. (....)

Now, while these screenshots seem to show that Tay has assimilated the internet's worst tendencies into its personality, it's not quite as straightforward as that. Searching through Tay's tweets (more than 96,000 of them!) we can see that many of the bot's nastiest utterances have simply been the result of copying users. If you tell Tay to "repeat after me," it will — allowing anybody to put words in the chatbot's mouth.

However, some of its weirder utterances have come out unprompted. The Guardian picked out a (now deleted) example when Tay was having an unremarkable conversation with one user (sample tweet: "new phone who dis?"), before it replied to the question "is Ricky Gervais an atheist?" by saying: "ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism."

But while it seems that some of the bad stuff Tay is being told is sinking in, it's not like the bot has a coherent ideology. In the span of 15 hours Tay referred to feminism as a "cult" and a "cancer," as well as noting "gender equality = feminism" and "i love feminism now".

Box 6

Barrat[65] is right in stating that it is not simply about infusing human values into an intelligent machine. He points out the fact that "Artificial General Intelligence" is also created to kill humans as, certainly in the US, national defense institutions are among the most active investors. Recently an international campaign was launched to the UN to outlaw "killer robots". A "peace robot" delivered a letter to UN diplomats demanding that robots not guided by human remote control, which could accidentally start wars or cause mass atrocities, should be outlawed by the same type of international treaty that bans chemical weapons.[66]

As we wish to find out how best to be in touch with robots, in view of the differences between human beings and robots, we can also learn from theology as Tilburg University Professor Paul van Geest explains (see box 6 below).

65 Barrat, J. (2013, p. 153)
66 McDonald, H. (2019)

"Theologians were the first to think about robots."
An interview with Paul van Geest, professor in Church History and History of Theology at Tilburg University

I was asked the question: Are theologians into robots? My immediate reply was: Yes! We were actually the first that thought about them. Almost two thousand years ago the rabbi Akiva was questioned about the issue who the better creator was: God or humankind? He answered: humans! Because God created corn but humans turn this into cookies. So it instantly became clear that humans perfect Creation.

There is an old Jewish tale, refined later on, about rabbi Löw from Prague in the 16th century, who observed the hard labor of people, even children, all day long. For that reason, he created—just like God created Adam—a puppet from clay that could serve humans and protect them. He called the puppet Golem. The precursor of the robot was born.

In the morning, a piece of paper was put into Golem's mouth, listing the things he should do, and at night, the rabbi removed the paper. The rabbi wanted Golem to behave like a human being and taught him to eat bread and read, but Golem could not make any distinctions and ate stones. He also tried to laugh and make fun like humans, but he just could not do it. Eventually, Golem wanted to become more and more like a human being; he wanted to be able to laugh and cry and be open-minded like a child. However, this was not granted to him, and he was frustrated with this; he began to destroy things. When men tried to catch him, he ran away.

In another version of the story, Golem wouldn't let go of the paper and refused to give it back. In this version, he also became bigger and bigger and rabbi Löw became afraid of him but was unable to remove the paper. When he finally succeeded in removing the paper, Golem became a big pile of clay and the rabbi was crushed under its weight. There are more versions; but the moral of this one is that people can succumb to what they have created themselves.

We are currently not living in paradise. Yet, some comparison with the creation story imposes itself. In this story, Adam and Eve were expelled from paradise because they were stubborn. Let us suppose that we create robots to bring us back to paradise. However, suppose that they, just like Golem did in the latest version, end up dominating us: that they become our bosses and become more powerful than we are. In that case, they will get us out of our comfort zone, and we would face a second expulsion from the Garden of Eden! [smiles]

At present, two main routes are distinguished for value alignment, a programming approach that is seen as a top-down approach and a bottom-up approach that aims at having the technology learn what we want it to by training and reinforcement. In the latter approach, TrusTee is more like a student than a slave, a view developed by Turing.[67]

Bostrom[68] describes what he calls the "value-loading problem," which could be seen as the problem of the top-down approach. He states that utility functions can be defined where an agent maximizes expected utility. Developing a machine that can compute a good approximation of the expected utility of the actions it has available is an AI problem, he argues. The second problem, however, is that if we wish the machine to contribute to, e.g., the value of happiness, it must first be defined in "in terms that appear in the AI's programming language and ultimately in primitives such as mathematical operators and addresses pointing to the contents of individual memory

67 Bouée, C.E. (2020, p. 6)
68 Bostrom, N. (2017, pp. 226-255)

registers (...) Identifying and codifying our own final goals is difficult because human goal representations are complex."[69]

He then goes on to identify strategies of value learning, which we have referred to before as the bottom-up approach. These strategies use the AI's capability to learn the values we wish it to pursue. In presenting an overview and evaluation of value-loading techniques, Bostrom starts with the conclusion that "It is not currently known how to transfer human values to a digital computer, even given human-level intelligence."[70] See box 7 below.

Box 7

Value-loading techniques identified by Bostrom

1. Explicit representation
2. Reinforcement learning
3. Value accretion
4. Motivational scaffolding
5. Value learning
6. Emulation modulation
7. Institution design

Techniques 1, 4 and 7 are evaluated as the most promising.[71] Nevertheless, Bostrom goes on to say, "If we knew how to solve the value-loading problem, we would confront a further problem: the problem of deciding which values to load. What (...) would we want a superintelligence to want?"[72] For this essay, this is the one question hitting the nail on the head.

Havens[73] is more optimistic or realistic about the value-loading problem: "(...) ironically enough, a lot of AI methodologies revolve around observing our ethical behavior as demonstrated by our actions. So they're already codifying our values, oftentimes without our direct input."

Wallach and Allen[74] state that values might unconsciously be built into technology and that it is not just a question of "engineering ethics." A great many engineers, companies,

69 Bostrom, N. (2017, p. 227)
70 Bostrom, N. (2017, p. 253)
71 Within the limitations of this essay we can not go deeper into details of these techniques. In literature on this topic, there is a wide array of other techniques being discussed (and criticized), e.g. the technique of Inverse Reinforcement Learning. See Arnold, T. et al. (2017)
72 Bostrom, N. (2017, p. 255)
73 Havens, J.C. (2016, p. xxi)
74 Wallach, W. & Allen, C. (2010)

and research centers are building the design of the new technologies and this complexity and division of labor implies that no one can actually have a complete prognosis of how the system will eventually interact and respond in an infinite number of situations.[75]

Etzioni and Etzioni[76] conclude that neither the top-down approach, in which ethical principles are programmed into the technology, nor the bottom-up approach works. In the latter approach, machines are expected to learn how to deliver ethical decisions through observation of human behavior in real situations, without learning any formal rules or being supplied with any specific moral philosophy.

As Wallach and Allen[77] state, we will also not easily succeed in turning values and ethics to "a logically principle or set of laws (...) given the complex intuitions people have about right and wrong." We should be careful with these top-down approaches although awareness of values and goals that we want technology to subscribe to is a condition sine qua non. In addition, yes, we do have to talk to designers, clients, producers, and users of the new technologies.

These observations urge us to understand that we have to move "beyond emphasizing the role of designers' values in shaping the operational morality of systems to providing the systems themselves with the capacity for explicit moral reasoning and decision making."[78] In other words, we should empower and educate our TrusTee robot. This, however, as has become apparent in this section, is easier said than done.

It also suggests that we have to accept that talking values with robots and AI is not something that we need to do in the initial design phase only. It will be a permanent conversation and communication. At Tilburg University, we are making many efforts in studying human-technology communication in both directions: making robots and AI understand what humans mean and want and vice versa. Natural language as well as psychophysiological and social signal processing and visual perception are key here.

Finally, if we could manage the value loading/learning process, we could decide to check whether our machines, like TrusTee, have fully understood what we mean and put them to a "Moral Turing Test" (MTT). The Turing test is quite well known. Turing wanted to avoid defining artificial intelligence through a set of ethical values. His idea was that a human evaluator would judge natural language conversations between a human and a machine. The evaluator is informed that one of the two conversation partners represents a machine.

If, as a result, the human evaluator cannot distinguish the machine from the human, the machine passes the test. The evaluation is not about the correctness of the answers, but merely about how closely the machine's answers resemble those a human would give. In the MTT, the human judge tries to determine, by asking questions, if he or she can reliably tell the machine's answers about morality from a human respondent.[79]

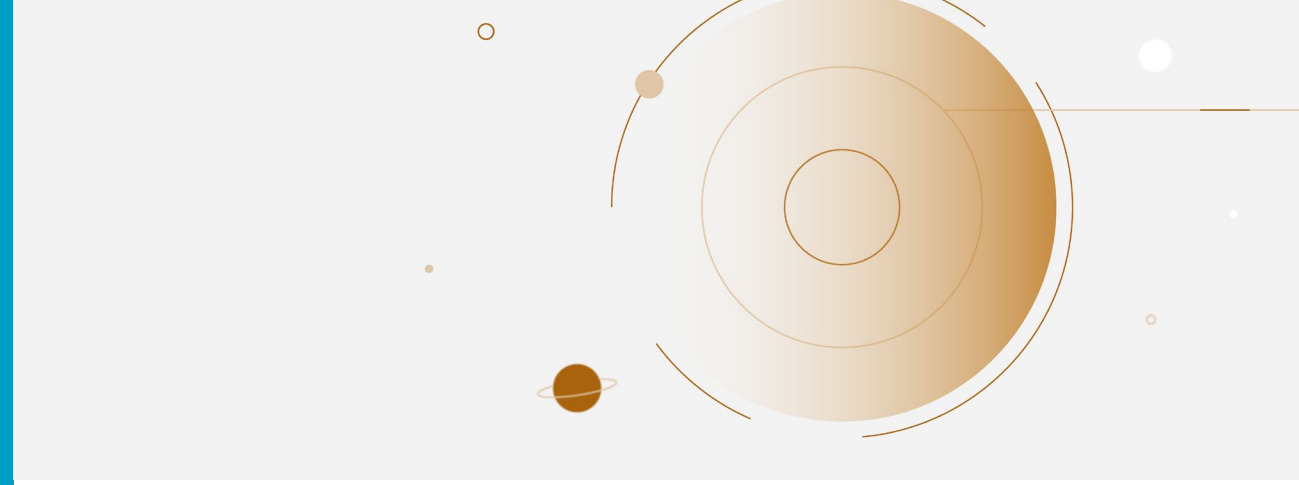
Arnold and Scheutz[80] strongly argue against such a moral test for autonomous systems. They raise concerns about the vulnerability of such a generally defined test to deception, inadequate reasoning, and inferior moral performance in view of a system's capabilities. These authors make a plea for "verification" that makes sense:

(...) we propose that a better concept for determining moral competence is design verification (...) a moral attribution must rely on more as an accountable, practical, socially implicated act of trust. To be accountable for a system's moral performance means going to as full a length as possible to verify its means of decision-making, not just judging ex post facto from a stated response or narrated action. Verification aims for transparent, accountable, and predictable accounts of the system's final responses to a morally charged context.[81]

Quite clearly: whether a robot or AI application can be deemed trusted and social needs to be verified, somehow, someway.[82]

75 See also Van de Poel, I. & Royakkers, L. (2011)
76 Etzioni, A., & Etzioni, O. (2017, p. 408)
77 Wallach, W. & Allen, C. (2010, p. 215)
78 Wallach, W. & Allen, C. (2010, p. 215)

79 Wallach, W. & Allen, C. (2010, pp. 206-207)
80 Arnold, T. & Scheutz, M. (2016)
81 Arnold, T. & Scheutz, M. (2016, p. 7)
82 See also Russell, S. et al. (2015, pp. 109-110)



6. Values vary and may conflict

Two complications arise in engaging technology in values. First, values vary and differ, as Dignum justly concludes:

Values are dependent on the socio-cultural context (...), and are often only implicit in deliberation processes, which means that methodologies are needed to elicit the values held by all the stakeholders, and to make these explicit can lead to better understanding and trust on artificial autonomous systems. That is, AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency. Responsible Artificial Intelligence is about human responsibility for the development of intelligent systems along fundamental human principles and values.[83]

The project European Value Studies, in which Tilburg University has been the leading partner, researches the empirical support for values across Europe. This survey shows both a great variety of value support and changes in supporting specific values. In the survey, the values are operationalized by statements that are presented to a sample of citizens of European countries. Values do not just vary across countries, but there are also similarities and differences among clusters of countries. A few examples indicate the scope of variety. One seminal example is the cross-national differences in social trust, i.e., trust in people different from you.[84] Comparative studies indicate that there is considerable variation in social trust (e.g., explained by a myriad of factors referring to national prosperity, good governance, and cultural legacy (most notably having a protestant tradition[85]). Also at the

83 Dignum, V. (2018, p. 1)
84 Uslaner, E.M. (2002)
85 Delhey, J. & Newton, K. (2005)

moral level, we see that there are differences across Europe in what we find acceptable, with a similar West-East and North-South pattern that distinguishes between morally permissive countries (where homosexuality, abortion, euthanasia, and divorce are more accepted) and rather restrained countries.[86] As a last example, to whom we show solidarity also results from the country we live in. Opinions about "who should get what, and why" from the welfare state flows from economic conditions countries are facing; in contexts of high unemployment, there is a demand for lower restrictions towards welfare claimants.[87]

Box 7

The European Values Study

The European Values Study is a large-scale, cross-national, and longitudinal survey research program on basic human values providing insight into the ideas, beliefs, preferences, attitudes, values, and opinions of citizens all over Europe. It is a unique research project on how Europeans think about life, family, work, religion, politics, and society.

The European Values Study started in 1981 when a thousand citizens in the European Member States of that time were interviewed using standardized questionnaires. Every nine years, the survey is repeated in an increasing number of countries. The fourth wave in 2008 covers no less than 47 European countries/regions, from Iceland to Armenia and from Portugal to Norway. In total, about 70,000 people in Europe were interviewed. The data collection of the fifth wave of this longitudinal research project was initiated in 2017. Presently, data of already 30 countries are publicly available, with more countries as well as an integrated longitudinal data file being published in the first half of 2020.

The project was published in a beautifully designed Atlas of European Value Studies. A rich academic literature is based on the surveys and numerous other works have made use of the findings. Data are freely available in several formats from the GESIS Data Archive and are compatible with the data from the World Values Survey. The database not only contains the data itself, but also full information on the used questionnaires (for all countries and languages that participated)—this information can be searched at the variable level. There is also an educational website.

The hard question here is whether, in building TrusTee, this Robot/AI should be flexible with regard to its socio-cultural context and should be able to make different decisions

86 Halman, L. & Van Ingen, E. (2015)
87 Van Oorschot, W. (2006)

in similar cases; or that we should develop a product range of different TrusTees for different socio-cultural markets. In Ian McEwan's recent novel *Machines like me*, both options are available. The main character, Charly, can choose for an Adam or Eve humanoid robot and can also pick a model tailor-made for a few Western or Arabic cultures. In addition, when installing the (Western) Adam, Charly (and his girlfriend) get the possibility of ticking a number of personal traits and values for Adam. Further below, we will discuss our suggestion of relying as much as possible on values that are democratically legitimized and have as much global scope as possible.[88]

A second complicating issue concerns value conflicts and the handling of these conflicts. We as human beings are certainly not unfamiliar with conflicting values, as individuals, but also at the collective and societal level. For instance, in legal conflicts and in public policy, values often compete for the attention and the judgment of decision-makers. Furthermore, some values, like those embodied in fundamental rights, have been attributed more weight than other values. There is not always a single and straight method of resolving value conflicts. As Thacher and Rein[89] explain:

E.g. policy makers do sometimes try to strike a "balance" among conflicting values, but they often avail themselves of other strategies as well: they cycle between values by emphasizing one value and then the other; they assign responsibilities for each value to different institutional structures; or they gather and consult a taxonomy of specific cases where similar conflicts arose.

In case robots and AI have to deal with value conflicts—which will happen—the situation might not appear very different, unless we agree and manage to feed technology with a clear hierarchy of values. Very often, the case of self-driving cars is referred to, where the car might have to decide between protecting the driver in the case of an accident or cause itself to crash in order to save the lives of a mother and several children in another car.

Admittedly, smart and intelligent technology may be faster and more accurate in processing huge amounts of data in calculating the broader costs and benefits of having some values prevailing over other values. However, this maximization of net gains might not meet the expectations of the humans involved and that of society in general. Researchers are trying hard to make progress in the field of value conflicts within autonomous actors. For example, Broersen et al.[90] propose a framework, called the BOID Architecture, where the actor has to prioritize Beliefs, Obligations, Intentions, and Desire, which can be done in different ways by different actors.

In building trusted and social robots and AI, handling value conflicts is a key issue, which relates to another issue, that of "value back-firing," that we address below.

88 McEwan, I. (2019)
89 Thacher, D. & Rein, M. (2004, p. 457)
90 Broersen, J. et al. (2001)

7. Identification of values: democratic legitimation

First, a strategy has to be chosen on how we can carefully identify the values that we can best take as points for departure in aligning with trustworthy technology. As Havens formulates the mission: "We need to codify our own values first to best program how artificial assistants, companions and algorithms will help us in the future." [91]

How are we going to do this, given the variation in values and the appreciation thereof? We feel that our best bet here is to start with democratically legitimized and shared values, agreed and laid down at the most universal and international levels and subscribed to by as many countries and parties as possible. Sharing core values is key to make a society function.

As Tony Wilkinson argues in *Capitalism and Human Values*[92], we need values, a framework of shared values, in order to ensure that efficient but sometimes remorseless economic systems, such as capitalism, lead to human flourishing rather than enslaving us. The importance of shared values is not only of paramount importance on a global level but certainly also at a local level, as illustrated in box 8 below.

91 Havens, J.C. (2016, p. xix)
92 Wilkinson, T. (2015)

Box 8

Sharing values at the local level

In 2016 Bart Somers, the Mayor of the City of Mechelen, Belgium (located between Antwerp and Brussels) received the World Mayor Prize[93]. He managed to transform a run-down, old industrial city with many social problems into a vibrant, attractive city with significant improvements in integration and social cohesion. Somers stresses the importance of sharing key values, guaranteeing every person's freedom. At the same time, he does not believe in relating shared values to a certain culture, to assimilation, where every citizen should belief, eat, do and like the same things. Instead, he radically started to fight poverty and, at the same time, re-established the rule of law.

Halman and Sieben observe a growing interest in values, also in view of political developments, and they expect some convergence but also resistance in the field:

In public and political discourses about a (dis)united Europe and its future development, the issue of values has come to the fore (...). Discussions about joining or leaving the European Union are not only economically inspired but center around the acceptance of Europe's core fundamental values as they are laid down in the Lisbon Treaty and EU's Charter of Fundamental Rights: human dignity, freedom, democracy, equality, the rule of law, and respect for human rights. The intensification of worldwide social relations, international trade, and flows of information and people will (...) lead to an increasing cosmopolitan outlook and ultimately to a homogenization of cultures. Consequently, the end of clear distinctive national identities and the gradual disappearance of cross-national differences in fundamental values fuel a cultural backlash of national and traditional values by those who feel threatened by these developments.[94]

Spijkers studies the evolution of global values and international law in relation to the founding, purposes, and policies of the United Nations (UN). He defines a global value as "an enduring, globally shared belief that a specific state of the world, which is possible, is socially preferable, from the perspective of the life of all human beings, to the opposite state of the world."[95]

93 See also in depth interview (in Dutch) in VNG Magazine.
94 Halman, L. & Sieben, I. (2019, pp. 1-2)
95 Spijkers, O. (2011, p. 9) This definition is inspired by Rokeach's definition of a value, which is also frequently referred to in psychological literature. Rokeach, M. (1973)

Looking back at the publications on value alignment (or value loading) and Harris' point about the science of values we discussed previously in this essay, this definition already provides an interesting perspective as a basic design principle for trusted/trustworthy robots and AI, at a very general level.

Spijkers then goes on to examine in depth four values that guide global decision-making, so values that do not specifically refer to individuals or communities, basing his research on the UN Charter and the resolutions of the UN General Assembly.

The values are the following: the value of peace and security, social progress and development, human dignity, and the self-determination of peoples.

The UN have produced two more contributions of paramount importance to a set of universal values. The Universal Declaration of Human Rights (UDHR) is a milestone document containing 30 fundamental human rights, to be protected universally. It was drafted by representatives with different legal and cultural backgrounds from all regions of the world and proclaimed by the United Nations General Assembly in Paris on December 10, 1948 as, in the words of the UN, a common standard of achievements for all peoples and all nations. Article 1 reads: "All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood."

On January 1, 2016, the 17 Sustainable Development Goals (SDGs) of the 2030 Agenda for Sustainable Development—adopted by world leaders at a 2015 UN Summit—officially entered into force. Although not legally binding, these goals are meant to apply to all countries, expecting to mobilize efforts "to end all forms of poverty, fight inequalities, and tackle climate change, while ensuring that no one is left behind." It is seen by the UN as "the blueprint to achieve a better and more sustainable future for all." The SDGs address global challenges, including poverty, inequality, climate, environmental degradation, prosperity, and peace and justice. The SDG's are displayed in box 9.

Box 9

The UN Sustainable Development Goals

At the pan-European level, The European Social Charter, established as a Council of Europe treaty is the point of reference. It guarantees fundamental social and economic rights similar to the European Convention on Human Rights, which addresses civil and political rights. The Charter deals with a broad scope of everyday human rights in the domains of employment, housing, health, education, social protection, and welfare, requiring that these rights can be enjoyed without discrimination. Most of the social rights in the EU Charter of Fundamental Rights are based on the relevant articles of the Charter, which means they are part of European Law.

Both the Council of Europe and the European Union are very active in the area of autonomous and intelligent technology, especially with regard to AI. The Council of Europe is a leading human rights organization with a membership of 47 countries that go beyond the borders of the European Union. For instance, the Russian Federation, Ukraine, and Turkey are among its members and the United States and some other countries have an observer status. The Council establishes Treaties/Conventions in various areas of rights and security.

Under the heading "Towards an ethical and responsible AI for human rights, rule of law and democracy," the Council has formed the ad hoc Committee on Artificial Intelligence (CAHAI), mandated by the Committee of Ministers to "examine the feasibility and potential elements on the basis of broad multi-stakeholder consultations, of a legal framework for the development, design and application of artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law."

As Marija Pejcinovic Buric, Secretary General of the Council of Europe, states on the Council's website the aim is to arrive at a global benchmark by a multi-stakeholder approach with other international organizations, civil society, businesses, and academia. Eventually a world-wide binding "alignment framework" may be in sight.[96]

The European Commission and the European Parliament are not only deeply involved in the regulation of AI and robotics but in its promotion as well. Europe as a continent is determined to get a solid position in the development of the technology for both economic and political reasons. At the same time, Europe wishes autonomous and intelligent technology to adhere to ethical standards. The European Commission has mapped out a European approach to Artificial Intelligence and Robotics.

On April 25, 2018, the European Commission issued the Communication Artificial Intelligence for Europe, Building Trust in Human-Centric Artificial Intelligence. It is stated that "The European AI strategy and the coordinated plan make clear that trust is a prerequisite to ensure a human-centric approach to AI: AI is not an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being.[97]

96 Jan Kleijssen, Director, Information Society and Action against Crime Directorate of the Council of Europe, personal communication.
97 European Commission (2018)

The EU High-Level Expert Group on Artificial Intelligence has drafted an elaborated set of guidelines for trustworthy AI, following a first draft that was put open to consultation.[98] Organizations can now fill out the assessment list for trustworthy artificial intelligence and find out how robust theirs is in practice. An online survey has been set up to collect feedback on the assessment list and will be open until December 1, 2019.[99] Best practice examples for assessing the trustworthiness of AI can also be shared through the European AI Alliance.

The set of guidelines is included in box 10

Box 10

High-Level Expert Group on Artificial Intelligence, Ethics Guideliness for Trustworthy AI

- Develop, deploy and use AI systems in a way that adheres to the ethical principles of: *respect for human autonomy, prevention of harm, fairness and explicability*. Acknowledge and address the potential tensions between these principles;
- Pay particular attention to situations involving more vulnerable groups such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of exclusion, and to situations which are characterised by asymmetries of power or information, such as between employers and workers, or between businesses and consumers;
- Acknowledge that, while bringing substantial benefits to individuals and society, AI systems also pose certain risks and may have a negative impact, including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.) Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk;
- Ensure that the development, deployment and use of AI systems meet the seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.
- Consider technical and non-technical methods to ensure the implementation of those requirements. Foster research and innovation to help assess AI systems and to further the achievement of the requirements; disseminate results and open questions to the wider public, and systematically train a new generation of experts in AI ethics.
- Communicate, in a clear and proactive manner, information to stakeholders about the AI system's capabilities and limitations, enabling realistic expectation setting, and about the manner in which the requirements are implemented. Be transparent about the fact that they are dealing with an AI system.
- Facilitate the traceability and auditability of AI systems, particularly in critical contexts or situations.
- Involve stakeholders throughout the AI system's life cycle. Foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
- Be mindful that there might be fundamental tensions between different principles and requirements. Continuously identify, evaluate, document and communicate these trade-offs and their solutions.
- Adopt a Trustworthy AI assessment list when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.
- Keep in mind that such an assessment list will never be exhaustive. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders in this.

Also in this expert report, the UN-principle of leaving no-one behind is reflected and made concrete in the development and design of technology.

Alignment frameworks for AI (and robotics) are not only proposed by formal, political institutions. No fewer than 126 manifests, declarations, or recommendations on responsible, responsive, trustworthy, good, humane, human-centric, or ethical AI have already been counted. In general, these manifests are not very specific.[100] Some come from scientists that worry about the direction things are taking; some come from industries or NGOs.

98 European Commission (2019)
99 You can register online to receive the survey link.

100 Jan Kleijssen, personal communication. See also De Jong, R. et al. (2019)

Many manifests issue negative concerns, starting-points, standards, rules, or regulations, i.e., indicating what AI and robots should *not* be or do. Other expressions contain positive expectations about what technology should *help* achieve.

All the values, rights, goals, and standards listed form a very interesting and relevant data set for research and the list is not complete by far. Think about the basic laws that countries have enacted, but also about more domain-specific conventions that have been ratified by a large share of countries, e.g., those of the International Labour Organization, including the right to collective bargaining and negotiations for representatives of workers, in the domain of employment.

Because all those value-rich laws and agreements have been established by a majority—of countries or parliaments—they cannot easily be altered or abolished; democratically legitimated values have the most gravitas, universality, and timelessness and are the most thoroughly justified. It is the democratic legitimation that makes them solid and a good starting-point for informing (to be) trusted technology.

The mission of the multidisciplinary Tilburg Institute for Law, Technology, and Society (TILT), founded 25 years ago, is to explore and understand the interplay between technology, regulation, and fundamental values/human rights; to study emerging technologies, their impact on the individual and society; and to assess the need for regulation of technologies. When it comes to the artificial agency of robots, the claim is that designers and regulators should look at the question of how to regulate robots' behavior in such a way that makes it compliant with legal norms.

The Dutch Rathenau Institute on research and dialogue relating to the societal aspects of science, innovation, and technology is also very active in this field and has relevant expertise for our purpose by defining new human rights in the light of the development of new technology.[101]

So how can we make sense of this large data on expressions of core values? TrusTee, our trusted robot, or more specifically AI can actually come to the help of researchers. The documents and texts where those values have been encrypted can be put into one database. We can use text analytics (or text mining) to cluster the values and make further analyses. There is no doubt that this is a time-consuming and tough job to do, but when it is done, we as human beings can finally tell robots and AI what we really value and consider worthwhile.

101 See in particular Van Est, R. & J.B.A. Gerritsen, with L. Kool (2017).

8. Value implementation and realization

Besides our values, there is one more issue we need to discuss with TrusTee. Surely, humans have values and we can identify them though this is not without complexity, as we saw. This issue concerns the implementation and realization of values. Values are not worth much if we do not succeed in achieving outcomes that result from these values or at least bring them forward. Just paying lip service to values does not get us anywhere.

Tilburg University colleagues Marjolijn Antheunis en Emmelyn Croes experiment with digital confession: Do we trust chatbots with our problems, secrets, weaknesses, and limitations? The tentative answer is that we do.[102] Without turning TrusTee into a priest, we have to provide intelligent technology that is being developed to work with and for us, with information on what we have achieved so far and what we see as obstacles. As all first-year students in sociology of law classes are taught: law in the books is not the same as law in action.

To give an example, one key value we have formulated as humans is equality, also between the sexes. Although progress has been made, there is still quite a pay gap, worldwide, between men and women, even within the same job with the same personal characteristics (age, education, et cetera). If we take universities as an example—women are still underrepresented in leadership positions in universities, including professorships.

102 One of the experiments took place at the 2019 edition of the Dutch music festival 'Lowlands'. The researchers controlled for alcohol consumption by means of alcohol testers, as people that have consumed alcohol possibly show inhibited behavior, which can explain why they share intimate secrets.

Informing intelligent technology about the obstacles in value implementation and realization is a mammoth task, but it is not impossible. Through a meta-analysis, we would need to take stock of all the fundamental and empirical conditions that apply to this and assess the state of the art regarding our key values. As in identifying values, artificial intelligence can help us in systematically collecting and reviewing the theoretical insights and empirical data from the best scientific papers (from the best journals with the highest rankings). What worked, what did not – what were and still are the critical conditions for effecting core human values and goals?

For example, with respect to the issue of equal treatment of men and women, it becomes clear that cultural barriers are still very strong, as women are strongly held responsible for the going on of family life and the education of children compared to their male partners. Negotiation skills are also mentioned as a cause of the gap, in addition to the HR assessment and evaluation systems that not always acknowledge that women on maternity leave cannot do research and write papers during those periods and are judged as being less productive. One finding here relates to the so-called pipeline problem. Research shows that "merely increasing the pool of qualified women has not led to a commensurate number of women rising to the top in academia. Women are still ending up in lower paid jobs, and they continue to earn less than men in comparable positions."[103] So, there is more going on.

Admittedly, as humans, we have our restrictions. We have to deal with what the French writer Camus has called *la condition humaine*, the human condition, which is hard to escape. Just like the mythological figure Sisyphus (in Camus' *The Myth of Sisyphus*), we sometimes have to keep pushing an enormous boulder up a hill only for it to roll down when it approaches the top, repeating this action time and again.

There are other notions too. In 1968, The American ecologist Garret Hardin wrote about the Tragedy of the Commons[104], arguing that individuals will always try to maximize their own gains even at the cost of the common good. He already indicated a typical human feature that would later on be called short-termism. For humans, not just for managers and corporations, it is difficult to balance the needs of both the long term and the short term.[105] We learned this the hard way during the recent financial crisis. Ideally, short-term actions and decisions translate into optimal long-term consequences, but in practice many actions we undertake favor the short-term and impair the long-term effectiveness.[106]

103 Monroe, K., & Chiu, W. (2010, p. 303)
104 Hardin, G. (1968)
105 See also Aspara, J. et al. (2014)
106 Marginson, D. & McAulay, L. (2008)

More recently, we have gained the insight from neuropsychology that short-term orientations have a basis in our brain structure and brain development. We have difficulties in delaying gratification or, in other words, controlling the impulse for immediate gratification. Within the midbrain, the limbic system constitutes both the reward center and the pleasure center, seeking instant pleasure. This system can be overridden by activity of the prefrontal cortex, historically a relatively new developed part of the human brain. The prefrontal cortex represents skills like planning, reasoning, learning from feedback, and focusing attention. It develops mainly during childhood, adolescence, and early adulthood. Children that are better able to control impulses will also be better at this a later stages in life.[107] The good news is that deferred gratification can be trained and learned, even by CEOs.[108]

These views appear fairly dystopian. However, progress has certainly been made in human history, as convincingly argued and documented by authors like John Norberg in Progress: *Ten Reasons to Look Forward to the Future*[109] and Hans Rosling in *Factfulness: Ten Reasons We're Wrong About The World - And Why Things Are Better Than You Think*.[110] We tend to underestimate, by ignoring facts and figures, what has actually been achieved, and we are on average much healthier, wealthier, and safer than at any point in history.

At the same time, we also have to concede that a huge part of the world population is still suffering and that new challenges have arrived which were systematically denied—such as climate change—or were not anticipated, e.g., the sharp increase of the forcibly displaced global population. According to UNHCR, by the end of 2018, almost 70.8 million individuals were forcibly displaced worldwide because of persecution, conflict, violence, or human rights violations, a record high. Very recently, Royal Dutch Shell Chairman of the Board of Directors Charles O. Holliday stated "the earth is dying" and announced that the company will move into hydrogen.[111]

Also in a well-developed and rich country as the Netherlands the number of working poor has been rising as of 1990. If the average levels of health, wellbeing, and participation are increasing, this is still a rather miserable result for those who are far below that level.

In view of our shortcomings and limitations, a school of thinking has emerged that is called "transhumanism," aiming at "engineering the human condition."[112]

107 Resnick, B. (2018)
108 See also Sitskoorn, M. (2017), which explains how one can use the plasticity of the brain to develop oneself.
109 Norberg, J. (2016)
110 Rosling, H. (2019)
111 It might be argued that if one concludes that the earth is dying, one could have determined at an earlier stage that the earth was already falling ill. Read more here (in Dutch)
112 Manzocco, R. (2019). On the first page of the book Nietzsche is cited: "You have evolved from worm to man, but much within you is still worm."

There are various interpretations of what transhumanism means. One is simply defeating individual death and making us live forever by applying science and technology. Maybe turning us all into robots (or half a robot) or providing "genetic doping," allowing us to be a better man and woman—human enhancement, i.e., more human than human. Another transhumanist direction is simply God Building, whereby God might be infinitely complex. Manzocco concludes: "*Now* this looks like the end-game of the Transhumanist endeavor: not simply to escape entropy, but to pursue an infinite level of complexity."[113] Transhumanism is no longer an obscure school of thought; various well-known writers of the past and present view themselves a transhumanist to some degree.

We can conclude that there is still much work to do in society and on the planet. We should also let TrusTee in on our flaws, weaknesses, struggles, and failures. This is what good friends do. We need all disciplines, the social sciences and humanities in particular, to identify, in addition to our values, the critical conditions of realizing and effectuating human values. However, we should still separate our roles and stir away from transhumanism.

113 Manzocco, R. (2019, p. 266)

9. Together we can do better

Working in tandem, human and robots can potentially do better than humans did before. We might actually benefit from a new form of co-creation between humankind and technology. There are great prospects—a new dawn, a new day—to make substantial progress in all sectors of life and industry: health care, sustainability, poverty, starvation, mobility, rule of law, democracy, citizenship, the work environment, finance, et cetera.[114] Box 11 shows how algorithms can contribute to the UN's sustainability goal of Zero Hunger in the world.

Box 11

Algorithms contributing to Zero Hunger in the world

Tilburg University and the Dutch Ministry of Foreign Affairs will invest 2.5 million euros for a period of 3 years to elevate the research into ending hunger and enabling Africa to feed itself. Tilburg University researchers have already helped to save the lives of millions from starvation through the smart application of data science. For instance by improving nutrition measures of food aid and placing more emphasis on local and regional agro and logistics.

Humanitarian workers of the UN World Food Programme in Yemen, Mozambique, Ethiopia, and Iraq are already using the Tilburg University algorithms to feed 15-20% more people affected by war or natural disaster with the same (and often limited) amount of money.

114 See Hudson, J. (2019).

Professors Hein Fleuren and Dick den Hertog of the Tilburg School of Economics and Management explain that the lab will be an open platform that will fast-track the collaboration between universities, government, NGOs, businesses, and local communities to rethink food and nutrition security world-wide. "Globally 1 billion people suffer from malnutrition, one third of the food produced gets lost or wasted, and 2 billion people are overweight. It is crystal clear that the world is not feeding itself in a healthy nor sustainable way. By initiating this project, we want to change the situation by enabling local farmers, communities, logistics service providers, and government to fight hunger."

The aim of building the Zero Hunger Lab to deliver innovative data science to achieve zero hunger by 2030 is in line with the Tilburg research vision of Science with a Soul. It is a prime example of Tilburg's Impact Program. The project underlines the Sustainable Development Goals of the UN.

If we make progress in adequate communicating with social robots via natural language processing, great opportunities emerge, as we can help people from different cultures to communicate. Tilburg University colleague Paul de Vogt was the first world-wide in experimenting with a social robot that offers a language program to toddlers from different cultures with different mother tongues. The interaction with the robot supports the children in learning a second language, a key condition in integrating and being successful in another culture and society.[115]

Getting the best out of intelligent technology and contributing to social and technological innovation works best in a transdisciplinary way, aiming at co-creation, with the involvement of all stakeholders (a broader concept than shareholders). Transdisciplinarity can create an ecosystem of open innovation with added value, where all the different expertise, knowledge, and resources can be brought together and deepened. This also leads to "inclusion by design," rather than trying to include stakeholders afterwards in the application phase. An initiative for co-creation taken by Tilburg University, Tilburg Municipality, and other stakeholders, such as schools, is described below in box 12.[116]

115 Willemsen, B. et al. (2018)

116 Another example of jointly exploring the challenges and potential of the new technology is the four months period 'Brabant Robot Challenge' where Tilburg University, Eindhoven University of Technology and the applied universities of Fontys and Advans match students from the four educational institutions to health care organizations with the aim of addressing key questions these institutions have with respect to robotization. These questions are not technological in the first place, but legal, ethical, economic, psychological and organizational.

Box 12

Innovation and co-creation – the example of Mind Labs

Mind Labs is a collaborative initiative between Fontys University of Applied Sciences (Journalism), intermediate vocational education (ROC Tilburg), media and publishing company *De Persgroep*, Tilburg Municipality, the Province of Noord-Brabant, and Tilburg University (Cognitive Science & Artificial Intelligence). Mind Labs operates in the domain of interactive technologies and behavior. It investigates human minds, artificial minds, and pushes an innovative mindset.

Mind Labs focuses on four research themes:

1. The Robotics & Avatars Lab develops virtual, intelligent avatars who can interact with humans.
2. The Serious Games Lab develops and investigates opportunities to bring fun and learning together.
3. The Natural Language Technologies Lab develops and tests computational linguistic algorithms to make computational sense out of human language.
4. The Virtual and Mixed Reality Lab investigates the intersection between our own world and simulated worlds.

Mind Labs aims for a collaboration between academic universities, universities of applied sciences, and intermediate vocational education. It builds on the idea that by emphasizing the uniqueness of these different knowledge institutions, collaborations become stronger. Mind Labs welcomes corporate partnerships and entrepreneurial minds, stimulating knowledge institutions to work alongside and together with companies. The iconic hotspot in the heart of the City of Tilburg makes Mind Labs a unique hub in the province of Noord-Brabant. Mind Labs is where minds, media, and technologies meet.

There is also a specific strategy for enhancing the impact of the application of the technology, that of certification. Creating TrusTee-certified robot and AI technology can literally produce good value because its users and clients will be assured that they can trust that this technology respects and promotes our human values and goals to the best extent possible. The differences between certified value-aligned and a bit less value-aligned may also be very subtle, but no less important.

Imagine that you are looking for a good care center for your ageing mother in ten or fifteen years' time. Intelligent technology will probably be all of over the place in all care centers available. By choosing a TrusTee-certified care center, you can trust that the technology in this center will serve to treat your mother to the highest standards based on human values and goals. Let us be very concrete: your mother attempts to

rise from her chair in the care center. A typical care robot might rush to you mother and lift her from the chair. The TrusTee-certified robot will decide otherwise and act from the principle of a patient's autonomy and self-dependency. Of course, TrusTee will help your mother but lets her get up by herself as much as she can and only will then start supporting her.

Certification of technology is a common strategy and there is no doubt that this will be applied to robots and AI to a larger extent, e.g., in terms of safety and quality standards. Technology might look good, even attractive, but has poor reliability.[117] The challenge here is to expand the method of certification to include human values as well, as described and discussed in this essay. We can build on many examples that already exist in the domains of food, clothing, and the environment. Again, the possibility of adequate verification is a necessary condition. If technology is granted the certificate "value-proof" by an authorized certification agency, it should actually work as such in practice. The method of assessing this should be robust and not prone to permanent changes.

These are ways to create (ro)bots and AI that collaborate and work with us. Empowered by intelligent, powerful, and trusted technology, we can, in the words of Yuval Noah Harari, set "A New Human Agenda."[118]

117 E.g. on the inaccuracy of current smartphone apps for measuring blood pressure see Jamaladin, H. et al. (2018), Raichle C.J. et al. (2018)
118 Harari, Y.N. (2017)

10. Value backfiring

There is something else to discuss. Suppose robots, AI and Artificial Moral Agents or however we wish to call them, become successfully infused and loaded with all the values and goals we as humankind cherish; an optimal situation of value-alignment. In addition, let's further suppose that these machines and forms of superintelligence are perfectly able and willing to decide and act according to these values and can help us to reach the good or even perfect society and save our planet, then we still have two related questions on the table.

The first question is about how we as humans will respond to these fully value-driven outcomes that the technology produces. When giving an example, we can borrow again from Ian McEwan's novel *Machines like me*[119]. In this book, the main character Charly, 32 years old and living in south London, buys an "Adam:" a robot from the first series of really human-like humanoids, which you can hardly distinguish from human beings. Adam is very able and earns a lot of money on the stock market, by just sitting in front of the computer. He outperforms Charly, who used to make a living that way. With those financial gains, Charly and his girlfriend are about to buy a great house in Notting Hill. Then Adam stays away for two days, and when Charly wakes up, Adam has arrived home and hands Charly a brown envelope with a modest amount of money. Seeing Charly puzzled, Adam explains that he visited a number of promising social projects in the neighborhood and decided that it makes much more sense to invest money in these projects, rather than buying a big and expensive house for just two young people.

For these situations, we propose to use the term "value backfiring". Adam, the robot, has acted perfectly on the human value of distributing wealth and investing in social cohesion. He also makes the perfect calculations, weighting the collective benefits

119 McEwan, I. (2019)

versus the individual gains for his owner. Nevertheless, the owner, i.e., the human, is utterly unhappy with this decision-making and value implementation process.

Evidently, these kinds of situations already occur in all domains of life, but so far, our judges have been humans: judges in court, police officers, other regulators, managers, doctors, family, friends, et cetera. Increasingly, robots and AI will hold up a mirror and we are not always going to like what we see. The reasons why we may be unhappy or disappointed are manifold, but amount to egocentrism, short-termism, or just, "no, not in my case," "not in my backyard," "not now," or just "don't feel like it." McEwan's book offers another far-reaching form of value backfiring when Charly's girlfriend is send to jail thanks to Adam's honesty and sincerity.

In the famous 2004 film *I, Robot*, loosely based on Asimov's writings, value backfiring is also the case. Eventually, Asimov added a "Zeroth Law" superseding even the First Law, which obliged robots to protect humanity as a whole, even if that means harming (or even killing) individual humans. In the movie, artificial intelligence supercomputer VIKI deems humans themselves to be the greatest danger to mankind (by causing environmental degradation, for instance) and justifies her murder of a scientist on the grounds of the Zeroth Law, which she autonomously invented. And, we are probably all familiar with the "computer says no" scenario, which is also dramatically portrayed in science fiction, originally in the book/movie *2001: A Space Odyssey*.[120]

The aforementioned relates to a second fundamental issue: can we as human beings live in a perfect world and do we really aspire to this? In *2001: A Space Odyssey*, an expert by the name of Floyd is flying from the earth to the moon, where meanwhile earthlings have set up a large basecamp and research center to check out an alarming situation of the possible presence of aliens. During the flight, the expert scans the electronic headlines of news from earth. Accidents, crimes, and man-made disasters are still the main concern. "Yet Floyd also wondered if this was altogether a bad thing; the newspapers of Utopia, he had long ago decided, would be terribly dull."

Evidently, as long as humans have not transformed into robots[121], which many people fear and some transhumanists hope, human imperfectness will not fully match artificial

120 Currently in The Netherlands, experiments are going on with an online assessment of the labor market position, income and competences of applicants for a mortgage. The reason is that in the past people's creditworthiness was mainly assessed by the existence (at the part of the applicant) of a permanent employment contract with an employer. Now the Dutch labor market is much more flexible and divers, with self-employed workers and various forms of temporary employment, the idea is that credibility should be measured on the basis of human capital rather than by contractual position (after all, a permanent contract no longer guarantees employment security, as businesses are restructuring, also because of technology). It is deemed possible that creditworthiness can be measured online in a few steps, but it is debated whether this means that people can be denied a mortgage so easily without the interference of a human, i.e. a representative of a bank, in a face-to-face conversation.

121 In (science) fiction, the idea that the next stage in the evolutionary line of mankind will be robots (or that we will blend into robots because gradually so much technology will be inserted in our body and brain – the cyborg) is quite popular, e.g. in Dan Brown's *Origin* from 2017, where a scientist discovered that this where we are going.

perfectionism. Moreover, we enter the "uncanny valley" when we experience something as strangely familiar, as Freud wrote in his 1919 essay Das *Unheimliche*, where he did not deal with robots but with the strangeness of dolls and waxworks.[122] There is abundant literature now on human-robot interaction and collaboration.

Besides, if we arrive in this perfect world, it might still not be altogether perfect, i.e., not for everyone and not without giving up our privacy.

Danish Member of Parliament Ida Auken has sketched the future of society in a blog as follows: see Box 13.

Box 13

Welcome to 2030

I own nothing, have no privacy, and life has never been better. All in all, it is a good life. Much better than the path we were on, where it became so clear that we could not continue with the same model of growth. We had all these terrible things happening: lifestyle diseases, climate change, the refugee crisis, environmental degradation, completely congested cities, water pollution, air pollution, social unrest and unemployment.

My biggest concern is all the people who do not live in our city. Those we lost on the way. Those who decided that it became too much, all this technology. Those who felt obsolete and useless when robots and AI took over big parts of our jobs. Those who got upset with the political system and turned against it. They live different kinds of lives outside of the city.

Professor of physics, Max Tegmark, talks about technology-driven Life 3.0 where both hardware and software can be designed. He expects unlimited prospects[123]:

To me, the most inspiring scientific discovery is that we've dramatically underestimated life's future potential. Our dreams and aspirations need not be limited to century-long life spans marred by disease, poverty and confusion. Rather, aided by technology, life has the potential to flourish for billions of years, not merely here in our Solar System, but also throughout a cosmos far more grand and inspiring than our ancestors imagined. Not even the sky is the limit.

122 Freud, S. (1919).
123 Tegmark, M. (2017)

What does this all mean for TrusTee? In any case, TrusTee would need to understand how the human mind and human emotions work. He/She should not merely be a champion in keeping up values, but also in "nudging"[124] us in valuable directions and give us a break whenever possible. Increasingly, very interesting applications of nudging can be found in the form of gamification.

Overall, humans should remain in control as much as possible unless others are harmed. TrusTee has no future as a value enforcer or terminator. Asimov's first and second law applies: do not injure a human being or, through inaction, allow a human being to come to harm; obey orders given by human beings except where such orders would conflict with the previous principle. As the EPSRC (Engineering and Physical Research Council, UK) puts it: Humans, not robots, are responsible agents, and robots should comply with existing laws.

It is important to distinguish between the aspects and roles of responsibility and accountability; two concepts that are often used interchangeably. Based on the widely applied RACI model, we can divide the following roles for man and machines. The machines could and should be held responsible (R) as the performers of tasks. Humans are the ones that are and remain accountable (A) for the correct completion and results of the task and delegating the jobs to the (R); they can also be considered the Approver, i.e., the final approving authority. Roles can change, accountability cannot.

Both humans, not being the (A) and machines, not doing the job (R), can act as consultants (C) in the sense that their opinions are valued and actively offered.

Finally, there are those that are informed (I), being kept up to date and brought up to speed on the progress and completion of the tasks and deliverables. Clearly, no one, neither humans nor machines, should be excluded from being part of (I).

124 See also Thaler, R.H. & Sunstein, C.R. (2009)

11. TrusTee and trusting ourselves: a research and impact agenda

In this essay, we have tried to sketch what it takes to create trusted technology, in the form of robots, softbots, and artificial intelligence. Our main argument is that it all starts and end with values in the design development process, as presented in section 2. The most trusted and social robot can never be built without us, as human beings, answering Tegmark's questions: "What does it mean to be human in the present day and age? For example, what is it that we really value about ourselves that makes us different from other life forms and machines?"[125] Wallach and Allen[126] write in their epilogue: "In writing this book, we have learned that the process of designing (ro)bots capable of distinguishing right from wrong reveals as much about human ethical decision-making as about AI."

Trust is the decisive factor. If we do not trust ourselves and others, we cannot build and trust (a) TrusTee. This is precisely the topic of Rachel Botsman's book *Who can you trust? How technology brought us together and why it can drive us apart,* in which she acknowledges that we still trust but not like we did before and that we must understand how trust is built, managed, lost, and repaired in the digital age.[127]

Based on trust, we can further engage into the design development process. As this process has a fundamental impact on society, on individuals, *and* on robots/AI, we have to deepen and speed up research along all steps of the design process. In this essay, we have identified some key points in doing this.

125 Tegmark, M. (2017, p. 82)
126 Wallach, W. & Allen, C. (2010, p. 215)
127 Botsman, R. (2018)

We need to:

- make sure that research and innovation and social sciences and humanities work hand in hand with technological research and innovation—it does not make any sense to make a hard distinction or even hierarchy among these sciences; they all need to be on board;

- operate in a transdisciplinary way, i.e., involve societal stakeholders in the research and development process, so that it becomes a co-creation approach with each stakeholder bringing in his/her own expertise, based on his/her distinct responsibilities;

- acknowledge that the new technology has tremendous moral significance and will, in its own way, decide and act morally, notwithstanding the accountability for this that human actors cannot shift to the technology;

- systematically identify and research the values that are central to humankind, including the values that we have or will develop with regard to the existence and functioning of non-humans. Norms are related to values and form a means to an end; in writing this essay, we conclude that two values stand out. First, above all, the pursuit of human well-being and second, as an organizational principle, inclusion, i.e., leaving no one behind;

- keep in mind that the appreciation of values differs across countries and societies and that values may conflict—trust being a value in itself. There are good reasons for working with those values that have strong democratic legitimation and a global basis of consensus;

- be aware of and systematically research the real-world critical conditions and obstacles that we as humans have experienced so far in implementing and realizing these values. This can only be done in an interdisciplinary way. These insights also need to be shared with (ro)bots and AI;

- understand that both top-down and bottom-up approaches of infusing technology with values and insights about value realization still have serious limits. Ultimately, we could perceive value alignment as a dialogue and learning process between humans and technology, shaping the operational morality of systems to providing the systems themselves with the capacity for explicit moral reasoning and decision-making;

- act on the starting-point that "together we can do better"—technology should be designed and implemented to protect the values we cherish and to improve our performance in the realization of these values;

- anticipate that robots and AI will also backfire our values to us, in the sense that the best decision, action, or consequence from the side of technology will not always be the decision, action, or consequence that suits us best. How are we going to deal with that?

Finally, currently there are manifold initiatives worldwide, targeting and investing in the development, application, and business models of AI and robots. This process looks like a contest: who will be the first, and who will be the strongest? Universities, cities, countries, continents, they all fear to be missing the boat and are competing for experts and resources. Often, the analysis is that China will be hard to compete with because of all the public investments in combination with fewer (ethical and regulatory) impediments; the US is also considered to be well positioned due to high private investments available.

Europe is speeding up its efforts because, although it is of sufficient size, it does not have the advantages and resources compared to China and the US. There is a fear that Europe might become a "digital colony."[128] Nevertheless, there is great value in taking the time to design and co-create trusted, responsible, and well-aligned technology. This will pay off in the end, not just financially but also in terms of broader value, acceptance, and impact.

In 1942, Julian Street published a book with the interesting title *Men, Machines and Morals* that is evidently not about the intelligent machines figuring in this essay. Street describes cases of how producers and suppliers of machines deal with morals, regarding both their customers and workers, in daily practice. The inspiring observation made here is that "good morals is good business."[129]

128 Read more here (in Dutch)
129 Street, J. (1942, p. 26)

12. About the authors

Prof. Ton Wilthagen is professor of Institutional and legal aspects of the labor market at Tilburg Law School. He is also one of the driving forces behind Tilburg University's Impact program, leading the theme 'Empowering the Resilient Society'. Ton is actively involved in social innovation. He is internationally known for developing the concept of flexicurity for the labor market. He also a member of the Dutch Coalition for Technology and Inclusion, initiator of the Brabant Robot Challenge and active within the Brainport Network, which he represents within the Dutch Technology Pact. For ten years, he taught a course on interdisciplinarity in the joint Research Master in Law of Tilburg University and the KU Leuven (Belgium).

Drs. Marieke Schoots is currently working as Impact Program Manager at Tilburg University. Within the Impact program, she is responsible for developing the theme Empowering the Resilient Society. She worked for several years as a member of the strategic staff of the Executive Board of TiU as a policy advisor in the field of valorization and knowledge transfer. From 2013 to 2017 she was seconded to the regional triple helix organization Midpoint Brabant. Within this secondment she set up the regional Social Innovation program. From 2013 to 2015 she was a board member of the European School for Social Innovation and from 2015 to 2017 managing director of the European Social Innovation festival Dear Future.

About Tilburg University Impact Program

Our society is more complex and diverse than ever and faces major challenges. Tilburg University's Impact Program originates in the involvement of our researchers in these challenges. In order to advance society, knowledge and innovation are needed, both in a societal and technological sense. Collaboration with knowledge institutions, social partners, businesses, and citizens is necessary. The Impact Program, together with partners, creates an impact on society. By this, we mean that scientific research contributes to solutions for complex issues in society, by building networks and communities to connect and collaborate on societal challenges. The Impact Team sets up programs and projects together with partners and raises funds to enable more research within three strategic themes: Empowering the Resilient Society, Enhancing Health and Wellbeing, and Creating Value from Data.

13. References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 7(3), 149–155. https://doi.org/10.1007/s10676-006-0004-4

Arnold, T. and Scheutz, M. (2018). The 'Big Red Button' Is Too Late: An Alternative Model for the Ethical Evaluation of AI Systems. *Ethics and Information Technology*, 20(1), 59–69. https://doi.org/10.1007/s10676-018-9447-7.

Arnold, T., & Scheutz, M. (2016). Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, 18(2), 103–115. https://doi.org/10.1007/s10676-016-9389-x

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). *Value Alignment or Misalignment - What Will Keep Systems Accountable?* Department of Computer Science, Tufts University. Medford, MA 02155. Retrieved from: https://hrilab.tufts.edu/publications/aaai17-alignment.pdf

Aspara, J., Pajunen, K., Tikkanen, H., & Tainio, R. (2014). Explaining corporate short-termism: self-reinforcing processes and biases among investors, the media and corporate managers. *Socio-Economic Review,* 12(4), 667–693. https://doi.org/10.1093/ser/mwu019

Auken, I. (2016, November 10). *Welcome to 2030: I own nothing, have no privacy and life has never been better.* Retrieved from: https://www.forbes.com/sites/worldeconomicforum/2016/11/10/shopping-i-cant-really-remember-what-that-is-or-how-differently-well-live-in-2030/#fcb6d4617350

Barrat, J. (2013). *Our Final Invention. Artificial intelligence and the end of the human era.* New York: Thomas Dunne Books.

Bostrom, N. (2017). *Superintelligence. Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Botsman, R. (2018). *Who can you trust? How technology brought us together and why it can drive us apart.* London: Penguin Books.

Bouée, C.E. (Expected 2020). *The Fall of the Human Empire: Memoirs of a Robot.* London: Bloomsbury Business.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z., & van der Torre, L. (2001). The BOID architecture: conflicts between beliefs, obligations, intentions and desires. *Proceedings of the Fifth International Conference on Autonomous Agents - AGENTS* '01, 9–16. https://doi.org/10.1145/375735.375766

Brown, D. (2017). *Origin*. New York: Doubleday.

Cheney, T. (2017, October 23). *AI? Robot? AMA?* Thinking about definitions. Retrieved from: https://thomascheneyblog.wordpress.com/2017/10/23/ai-robot-ama-thinking-about-definitions/

Chui, M., Manyika, J. & Miremadi, M. (2016, July). Where machines could replace humans and where they can't (yet). *McKinsey Quarterly*. Retrieved from: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet

Clifford, C. (2017, July 17) *Elon Musk: Robots will be able to do everything better than us.* Retrieved from: https://www.cnbc.com/2017/07/17/elon-musk-robots-will-be-able-to-do-everything-better-than-us.html

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology, 14*(1), 53–60. https://doi.org/10.1007/s10676-011-9279-1

Damasio, A., Christen, J., Changeux, J.P. and Singer, W. (2005). *The Neurobiological Grounding of Human Values*. Berlin: Springer.

Danaher, J. (2019). *Public debate 'Welcoming Robots into the Moral Circle'*. Tilburg Institute for Law, Technology and Society (TILT) on 24 September 2019.

Danaher, J. (2019). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*. 1-27 https://doi.org/10.1007/s11948-019-00119-x

de Jong, R., Kool, L. & van Est, R. (2019, March 19). *This is how we put AI into practice based on European Values*. Rathenau Instituut. Retrieved from: https://www.rathenau.nl/en/digital-society/how-we-put-ai-practice-based-european-values

Delaney, K.J. (2017, February 17). The robot that takesyour job should pay taxes, says Bill Gates. *Quartz*. Retrieved from: https://qz.com/911968/bill-gates-the-robot-that-takes-your-job-should-pay-taxes/

Delhey, J., & Newton, K. (2005). Predicting Cross-National Levels of Social Trust: Global Pattern or Nordic Exceptionalism. *European Sociological Review*, 21(4), pp. 311-327

Dignum, V. (2018). *Ethics in artificial intelligence: introduction to the special issue. Ethics and Information Technology*, 20(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way.* New York: Springer International Publishing.

Dolling, L. (2013, November 5). Humanities and Science Must Work Together. *The New York Times.* Retrieved from: https://www.nytimes.com/roomfordebate/2013/11/04/the-fate-of-the-humanities/humanities-and-science-must-work-together

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics, 21*(4), 403–418. https://doi.org/10.1007/s10892-017-9252-2

European Commission (2018). Communication from the commission to the European Parliament, the European Council, the Council, The European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe. *COM(2018) 237 final*. Retrieved from: https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF

European Commission (2019, April 8). *Ethics Guidelines for Trustworthy AI*. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Ford, M. (2016). *Rise of the Robots. Technology and the threat of a jobless future.* New York: Basic Books.

Freud, S. (1919). *The "Uncanny".* First published in Imago, Bd. V., 1919 [Translated by Alix Strachey]

Fry, H. (2018). *Hello World: How to be Human in the Age of Algorithms.* London: W.W. Norton & Company.

Gips, J. (2011). Towards the Ethical Robot. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics (pp. 244–253)*. https://doi.org/10.1017/CBO9780511978036.019

Glaser, L.B. (2016, March 18). How technology and humanities intersect. *Cornell Chronicle*. Retrieved from: https://as.cornell.edu/news/big-ideas-how-technology-and-humanities-intersect

Gunkel, D.J. (2018). *Robot Rights*. Cambridge: The MIT Press.

Gunnoo, H.A. (2019, June 1). Asimov's Laws of Robotics and Why AI May nit Abide by Them. *Towards Data Science*. Retrieved from: https://towardsdatascience.com/asimovslaws-of-robotics-and-why-ai-may-not-abide-by-them-e6da09f8c754

Halman, L. and Sieben, I. (2019). *Values*. Tilburg University. Under Review.

Halman, L., & van Ingen, E. (2015). Secularization and Changing Moral Views: European Trends in Church Attendance and Views on Homosexuality, Divorce, Abortion, and Euthanasia. *European Sociological Review,* 31(5), pp. 616-627.

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can You Trust Your Robot? *Ergonomics in Design: The Quarterly of Human Factors Applications,* 19(3), 24–29. https://doi.org/10.1177/1064804611415045

Harari, Y.N. (2017). *Homo Deus. A Brief History of Tomorrow.* London: Vintage.

Hardin, G. (1968). The Tragedy of the Commons. *Science, New Series,* 162(3859), 1243–1248.

Harris, S. (2012) *The Moral Landscape. How science can determine values.* London: Transworld Publishers.

Harwell, D. (2019, October 2019). A face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post.* Retrieved from: https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Havens, J.C. (2016). *Heartificial Intelligence*. Embracing our Humanity to Maximize Machines. New York: Jeremy P Tarcher/Peguin.

Hudson, J. (2019). *The Robot Revolution*. Understanding the Social and Economic Impact. Cheltenham: Edward Elgar.

Inglehart, F. (2019). *Cultural Evolution People's Motivations are Changing*, and Reshaping the World. Cambridge: Cambridge University Press.

Jamaladin, H., van de Belt, T. H., Luijpers, L. C., de Graaff, F. R., Bredie, S. J., Roeleveld, N., & van Gelder, M. M. (2018). Mobile Apps for Blood Pressure Monitoring: Systematic Search in App Stores and Content Analysis. *JMIR MHealth and UHealth,* 6(11), e187. https://doi.org/10.2196/mhealth.9888

Leenes, R., & Lucivero, F. (2014). Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design. *Law, Innovation and Technology,* 6(2), 193–220. https://doi.org/10.5235/17579961.6.2.193

Maio, G.R. (2017). *The Psychology of Human Values.* Oxon: Routledge.

Manzocco, R. (2019). *Transhumanism. Engineering the Human Condition*. Cham: Springer.

Marginson, D. & McAulay, L. (2008). Exploring the debate on short-termism: a theoretical and empirical analysis. *Strategic Management Journal*, 29(3), 273–292. https://doi.org/10.1002/smj.657

McDonald, H. (2019, October 21). Campaign to sop 'killer robot' takes peace mascot to UN. *The Guardian.* Retrieved from: https://www.theguardian.com/science/2019/oct/21/campaign-to-stop-killer-robots-takes-peace-mascot-to-un

McEwan, I. (2019). *Machines Like Me*. London: Jonathan Cape.

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, 49(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Monroe, K. R., & Chiu, W. F. (2010). Gender Equality in the Academy: The Pipeline Problem. *PS: Political Science & Politics,* 43(02), 303–308. https://doi.org/10.1017/S104909651000017X

Norberg, J. (2016). *Progress: Ten Reasons to Look Forward to the Future.* London: Oneworld Publications.

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence,* 26(3), 303–315. https://doi.org/10.1080/0952813X.2014.895111

O'Neil, C. (2016). *Weapons of Math Destruction*. New York: Crown Books.

Parsons, T. (1991). *The Social System*. London: Routledge & Kegan Paul (First published in 1951. New preface by Bryan S.Turner).

PILP (The Public Interest Ligitation Project) (2015, December 11). *Profiling and SyRI*. Retrieved from: https://pilpnjcm.nl/en/dossiers/profiling-and-syri/

Prinster, R. (2017, July 3). *Humanities, Social Sciences Have a Role to Play in a World Dominated by Technology.* Retrieved from: https://www.insightintodiversity.com/humanities-social-sciences-have-a-role-to-play-in-a-world-dominated-by-technology/

Rahwan, I. (2018). Society-in-the-Loop: Programming the Algorithmic Social Contract. *Ethics and Information Technology,* 20(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Raichle, C. J., Eckstein, J., Lapaire, O., Leonardi, L., Brasier, N., Vischer, A. S., & Burkard, T. (2018). Performance of a Blood Pressure Smartphone App in Pregnant Women: The iPARR Trial (iPhone App Compared With Standard RR Measurement). *Hypertension,* 71(6), 1164–1169. https://doi.org/10.1161/HYPERTENSIONAHA.117.10647

Resnick, B. (2018, June 6). The "marshmallow test" said patience was a key to success. A new replication tells us s'more. *Vox.* Retrieved from: https://www.vox.com/science-and-health/2018/6/6/17413000/marshmallow-test-replication-mischel-psychology

Rokeach, M. (1973). *The Nature of Human Values.* New York: The Free Press.

Rosling, H., Rosling, O. and Rosling Rönnlund, A. (2018). Factfulness: *Ten Reasons We're Wrong About the World - And Why Things Are Better Than You Think.* New York: Flatiron Books.

Russell, S., Dewey, D., & Tegmark, M. (2015). *Research Priorities for Robust and Beneficial Artificial Intelligence*. AI Magazine, 36(4), 105. https://doi.org/10.1609/aimag.v36i4.2577

Sacasas, L.M. (2011, August 25). *Kranzbergs Six Laws of Technology, A Metaphor, and a Story.* Retrieved from: https://thefrailestthing.com/2011/08/25/kranzbergs-six-laws-of-technology-a-metaphor-and-a-story/

Sitskoorn, M. (2017). *Train Your CEO Brain: And Become Your Best Self.* Deventer: Vakmedianet Management.

Spijkers, O. (2008). The United Nations and the Evolution of Global Values. *Human Rights Research Series.* Leiden Unviersity, Leiden, The Netherlands. Retrieved from: https://openaccess.leidenuniv.nl/handle/1887/17926

Street, J. (1942). *Men, Machines and Morals.* Granville: Denison University Press.

Sunstein, C. and Thaler, R. (2008). *Nudge.* New Haven: Yale University Press.

Tegmark, M. (2017). *Life 3.0. Being human in the age of Artificial Intelligence*. London: Penguin Random House UK.

Teubner, G. (2006). Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law. *Journal of Law and Society,* 33(4), 497–521.

Thacher, D., & Rein, M. (2004). Managing Value Conflict in Public Policy. *Governance,* 17(4), 457–486. https://doi.org/10.1111/j.0952-1895.2004.00254.x

Tilburg University (2019). Scientific Strengths. Retrieved from: https://www.tilburguniversity.edu/sites/tiu/files/download/scientific-strengths.pdf?utm_source=short-url

Tjong, Tjin Tai, E., Mak, V. & Berlee, A. (2018). *Liability for (semi-)autonomous systems.* Research handbook in data science and law. Cheltenham: Edward Elgar Publishing, pp. 55-82

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series,* 59(236), 433–460.

Uslaner, E.M. (2002). *The Moral Foundations of Trust.* Cambridge: Cambridge University Press.

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology,* 20(1), 27–40. https://doi.org/10.1007/s10676-017-9440-6

Van de Poel, I. and Royakkers, L. (2011). *Ethics, Technology and Engineering.* An Introduction. Chichester: Wiley-Blackwell.

Van den Herik, H.J. (2016). *Intuitie valt te programmeren*. Tilburg University. Retrieved from: https://www.universiteitleiden.nl/binaries/content/assets/science/lcds/20160202_oratie_web_vdn_herik_nl.pdf

Van Est, R. & Gerritsen, J. (2017). *Human rights in the robot age. Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality.* Rathenau Instituut. Retreived from: https://www.rathenau.nl/sites/default/files/2018-02/Human%20Rights%20in%20the%20Robot%20Age-Rathenau%20Instituut-2017.pdf

Van Oorschot, W. (2006). Making the Difference in Social Europe: Deservingness Perceptions among Citizens of European Welfare States. *Journal of European Social Policy,* 16(1), pp. 23-42.

Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics*, 25(3), 719–735. https://doi.org/10.1007/s11948-018-0030-8

Vincent, J. (2016, March 24). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. Retrieved from: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Wallach, W. and Allen, C. (2010). *Moral Machines. Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

Wallach, W., Franklin, S., & Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. *Topics in Cognitive Science,* 2(3), 454–485. https://doi.org/10.1111/j.1756-8765.2010.01095.x

Welzel, C., Inglehart, R., & Kligemann, H.-D. (2003). The theory of human development: A cross-cultural analysis. *European Journal of Political Research*, 42(3), 341–379. https://doi.org/10.1111/1475-6765.00086

Wilkinson, T. (2015). *Capitalism and Human Values*. Upton Pyne: Imprint Academic.

Willemsen, B., de Wit, J., Krahmer, E., de Haas, M., & Vogt, P. (2018). Context-sensitive Natural Language Generation for robot-assisted second language tutoring. *Proceedings of the Workshop on NLG for Human–Robot Interaction,* 1–7. https://doi.org/10.18653/v1/W18-6901

Wilthagen, T., Aarts, E., & Valcke, P. (2018). *Time for interdisciplinarity: An essay on the added value of collaboration for science, university, and society.* Tilburg: Tilburg University. Retrieved from: https://pure.uvt.nl/ws/portalfiles/portal/28373248/180464_Essay_Tijd_voor_Interdisciplinariteit_EN_finalproof_1_.pdf

Wilthagen, T., Denollet, J. & den Hertog, D. (2017). *Advancing Society in a Digital Era. Science with a Soul*. Tilburg University. Retrieved from: https://www.tilburguniversity.edu/sites/tiu/files/download/UVT0140_ESSAY%20IMPACTPROGRAMMA%202017_EN_2.pdf

For further information on specific scientific strenghts, please click on the title.
To return to the design development circle on page 13, click here.

Colofon

www.tilburguniversity.edu/impact
impact@tilburguniversity.edu
+31 134 664 512
Publication date November 2019

Production and Editing: Riet Bettonviel, Yvonne van Bruggen, Nina Karabetyan and Marieke Schoots

Printing: Canon Nederland N.V.

ISBN: 978-94-6167-409-8

TILBURG UNIVERSITY

Warandelaan 2
5037 AB Tilburg
T 0031(0)134669111
www.tilburguniversity.edu

www.facebook.com/TilburgUniversity
twitter.com/TilburgU

TILBURG ◆ UNIVERSITY