A Renewed Standpoint on More Ethical Technologies

Racism, blood shed, speciesism, climate change… Do you want a more ethical world? Dr. Thilo Hagendorff, researcher at the University of Tuebingen, argues you should give up ingroup and outgroup thinking patterns and develop instead an unconditional compassion. Do you want more ethical AI and technologies? Forget the dominant principle-based approach and adopt the virtue-based one. You should also start working on implementing an ethical working climate.

Before we begin, can you quickly explain your expertise which covers AI ethics, media ethics and technology ethics?

These are all vague terms, and there is a lot of overlapping between them. I have a background in philosophy, but I consider now to be alienated from this field, because I find many philosophical discourses to be somewhat obsolete. I prefer to conduct transdisciplinary research including sociology, science and technology studies and ethics. When I identify interesting research questions, I figure out if I can work on them on my own or if I need the expertise of other researchers. For example, I needed help to gather empirical data and conduct the recent study I published on the industry involvement in Machine Learning (ML) research[1]. This study analyzes empirically the ties between public and private institutions when it comes to ML research. Our material consisted of the articles of the three main ML conferences for the last five years. I hope a lot of people will read it so that future arguments can be more differentiated.

Faced with the plethora of ethical guidelines on AI, how can we verify that current developments really are more ethical?

This is a tough question, and I definitely don't have a definite answer. I think there are two sides of the same story that can be told: the pessimist and the optimistic views.

The pessimist view would argue the current AI developments are less ethical. Indeed, a lot of ethic washing is going on right now. Many companies are sending out strong and continuous signals to the public and to the legislator that they are handling AI ethics, so as to prevent legally binding regulations. Their message is "We have an ethics committee, there is no need for regulation, the influence of *soft law* is sufficient."

On the other hand, the optimistic view would emphasize that the current abundance of AI ethics work has the merit to raise awareness among companies and the population about the importance of values in AI, and even more broadly in technology.

Which one of these views hold true? Perhaps both at the same time.

How can ethical guidelines be concretely translated into technical developments?

[1] Hagendorff, Thilo, and Kristof Meding. "The Big Picture: Ethical Considerations and Statistical Analysis of Industry Involvement in Machine Learning Research." *arXiv preprint arXiv:2006.04541* (2020). https://arxiv.org/pdf/2006.04541v1.pdf

In a paper I published earlier this year[2], I compared 20 AI ethical guideline to analyze which values are most important, which others are underrepresented and which ones are omitted.

All these guidelines try to differentiate very abstract principles that are hardly operational. In addition, each principle can raise different approaches, take the privacy principle: one guideline will promote "privacy by design," another a checklist to guarantee minimal privacy, and so forth. This dominant deontological approach aims at a practitioner's adhesion to these principles and rules. However, empirical studies show that it does not change their daily practice. Therefore, I doubt this now dominant approach is promising. When the latter is based on values which are goals and immaterial mental concepts, we should consider another approach.

This other approach can be based on virtues which are dispositions that each individual can adopt and that affect one's actions, such as the disposition for care. It's easy to postulate principles for technologies and aim at socially acceptable goals, but it's quite hard to educate practitioners and strengthen virtues, so that, for instance, their actions tend to be more prone to justice. Apart from too few academic discussions (read Deborah Johnson on whether ethics can be taught to engineers[3], and this interesting study[4] on the framework needed for practitioners to adapt their day-to-day routine), most of the energy is spent on the principle-based approach, developing sets of principles and then transitioning from principles to practise, when we should try the virtues-based alternative.

Can practitioners with good virtues alone make ethical AI?

Principles alone cannot guarantee ethical AI. We need virtues, but these are personal dispositions. We therefore also need to implement working climates where ethical decisions are not sanctioned but rewarded. To understand these challenges, I recommend this is one great 2010 meta-study[5] that drew from over 30 years of research to determine the factors for unethical choice covering individual dispositions, moral issues and organizational environment (egoistic or ethical). Keeping in mind the different factors is key to understanding what is at stake, and that not everything comes down to individual dispositions, but several further factors like, for instance, an ethical working climate.

What is the basis for thinking about the ethics of non-humans?

Right now, we are having a completely anthropocentric view on these topics. People are having a moral status but not animals nor plants. At the same time, scientific discourses like new materialism or actor network theory are saying technical artifacts can act in our society,

[2] Hagendorff, Thilo. "The ethics of Ai ethics: An evaluation of guidelines." Minds and Machines (2020): 1–22.
[3] Can engineering ethics be taught? Deborah G. Johnson (2017) https://www.nae.edu/19582/Bridge/168631/168649.aspx
[4] Morley, Jessica; Floridi, Luciano; Kinsey, Libby; Elhalal, Anat (2019): From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices. In arXiv, pp. 1–21. https://arxiv.org/ftp/arxiv/papers/1905/1905.06876.pdf
[5] Kish-Gephart, Jennifer J.; Harrison, David A.; Treviño, Linda Klebe (2010): Bad apples, bad cases, and bad barrels. Meta-analytic evidence about sources of unethical decisions at work. In The Journal of applied psychology 95 (1), pp. 1–31. https://www.researchgate.net/publication/41087509_Bad_Apples_Bad_Cases_and_Bad_Barrels_Meta-Analytic_Evidence_About_Sources_of_Unethical_Decisions_at_Work

thus having a kind of agency. This is counterintuitive. The intuition would grant animals with agency, since they are closer to us!

The problem lies in setting ingroups and outgroups in our minds. People draw a circle around them: they include humans inside, when plants, animals and technical beings stay outside. Right now, we see very strong ingroup orientations. Based on studies from psychology, one can measure an individual's "social dominance orientation,"[6] a trait which evaluates the tendency to think in terms of in and out patterns and hierarchy in groups. This generates *speciesism* as well as racist and sexist discrimination.

From a pragmatic perspective, boundaries are necessary for human beings to learn and use terms when we try to grasp reality and thus necessary to understand the world. Moreover, people are prone to think in terms of in- and out-groups: it is inherited from evolution.

Nevertheless, I argue we need a more holistic approach to various topics like machine behaviour, climate change, migration, racism, factory farming and the like. It's really important to have an unconditional compassion toward members of outgroups. Bruno Latour, in his theory of the Parliament of Things, demanded a politics of artifacts. Even among human beings, Jacques Derrida denounced with his deconstructivist approach the numerous artificial differentiations that we do and that are, radically speaking, at the root of a lot of blood shed.

These fascinating discussions sadly remain mostly academic, there's a long way to go to reach mass media and the general public. Yet we must include nature and animals in our ethics, the fate of the planet is at stake.

Is it a good idea to also include robots? I don't know. People are compassionate with social robots, studies have shown that the same areas in the brains are activated when we see people and social robots hurt. Maybe we can exploit this sense of compassion.

Interview by Lauriane Gorce, Scientific Director of the Institut de la technologie pour l'humain—Montréal.

[6] Wikipedia definition: "SDO is conceptualized under social dominance theory as a measure of individual differences in levels of group-based discrimination; that is, it is a measure of an individual's preference for hierarchy within any social system and the domination over lower-status groups. It is a predisposition toward anti-egalitarianism within and between groups."