Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical

Luciano Floridi [1,2]

1 Introduction

It has taken a very long time,[1] but today, the debate on the ethical impact and implications of digital technologies has reached the front pages of newspapers. This is understandable: digital technologies—from web-based services to Artificial Intelligence (AI) solutions—increasingly affect the daily lives of billions of people, so there are many hopes but also concerns about their design, development, and deployment (Cath et al. 2018).

After more than half a century of academic research,[2] the recent public reaction has been a flourishing of initiatives to establish *what* principles, guidelines, codes, or frameworks can ethically guide digital innovation, particularly in AI, to benefit humanity and the whole environment. This is a positive development that shows awareness of the importance of the topic and interest in tackling it systematically. Yet, it is time that debate evolves from the *what* to the *how*: not just *what* ethics is needed but also *how* ethics can be effectively and successfully applied and implemented in order to make a positive difference. For example, the European *Ethics Guidelines for Trustworthy AI*[3] establish a benchmark for what may or may not qualify as ethically good AI in the EU. Their publication is currently being followed by practical efforts of testing, application, and implementation.

The move from a first, more theoretical *what* chapter, to a second, more practical *how* chapter, so to speak, is reasonable and commendable. However, in translating principles into practices, even the best efforts may be undermined by some unethical

[1] See (Floridi 2015) for references.

[2] In the ethics of AI, see for example (Wiener 1960; Samuel 1960).

[3] See (European Commission 8 April 2019), published by the High-Level Expert Group (HLEG) on Artificial Intelligence (AI) appointed by the European Commission (disclosure: I am a member of the HLEG).

✉ Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

[2] The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

risks. In this article, I wish to highlight five of them. We shall see that they are more clusters than individual risks, and there may be other clusters as well, but these five are the ones already encountered or foreseeable in the international debate about digital ethics.[4] Here is the list: (1) *ethics shopping*; (2) *ethics bluewashing*; (3) *ethics lobbying*; (4) *ethics dumping*; and (5) *ethics shirking*. They are the five "ethics gerunds", to borrow Josh Cowls' apt label, who also suggested to consider the first three more "distractive" and the last two more "destructive" problems. Let us consider each of them in some detail.

2 Ethics Shopping

A very large number of ethical principles, codes, guidelines, or frameworks have been proposed over the past few years. There are currently more than 70 recommendations, published in the last 2 years, just about the ethics of AI (Algorithm Watch 9 April 2019, Winfield 18 April Winfield 2019). This mushrooming of documents is generating inconsistency and confusion among stakeholders regarding which one may be preferable. It also puts pressure on private and public actors—that design, develop, or deploy digital solutions—to produce their own declarations for fear of appearing to be left behind, thus further contributing to the redundancy of information. In this case, the main, unethical risk is that all this hyperactivity creates a "market of principles and values", where private and public actors may shop for the kind of ethics that is best retrofitted to justify their current behaviours, rather than revising their behaviours to make them consistent with a socially accepted ethical framework (Floridi and Lord Clement-Jones 20 March Floridi and Clement-Jones 2019). Here is a more compact definition:

Digital ethics shopping = $_{\text{def.}}$ the malpractice of choosing, adapting, or revising ("mixing and matching") ethical principles, guidelines, codes, frameworks, or other similar standards (especially but not only in the ethics of AI), from a variety of available offers, in order to retrofit some pre-existing behaviours (choices, processes, strategies, etc.), and hence justify them a posteriori, instead of implementing or improving new behaviours by benchmarking them against public, ethical standards.

Admittedly, in a recent meta-analysis, we showed that much of the diversity "in the ethics market" is apparent and more due to wording and vocabulary rather than actual content (Floridi et al. 2018, Floridi and Cowls forthcoming). However, the potential risk of "mixing and matching" the list of ethical principles one prefers remains real, because semantic looseness and redundancy enable interpretative relativism. Ethics shopping then causes incompatibility of standards (it is hard to understand whether two companies follow the same ethical principles in developing AI solutions, for example), and with that a lower chance of comparison, competition, and accountability.

[4] See for example the debates about (a) the "Human Rights Impact Assessment of Facebook in Myanmar" published by the Business for Social Responsibility, https://www.bsr.org/en/our-insights/blog-view/facebook-in-myanmar-human-rights-impact-assessment; (b) the closure of Google's Advanced Technology External Advisory Council https://blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/; and (c) the Ethics guidelines for trustworthy AI, published by the High-Level Expert Group of the European Commission https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (disclosure: I was a member of the former, and still am a member of the latter).

The strategy to deal with digital ethics shopping is to establish clear, shared, and publicly accepted ethical standards. This is why I recently argued (Floridi 2019) that the publication of the *Ethics Guidelines for Trustworthy AI* is a significant improvement, given that it is the closest thing available in the European Union (EU) to a comprehensive, authoritative, and public standard of what may count as socially good AI.[5] Now that the *Guidelines* are available, the malpractice of digital ethics shopping should be at least more obvious if not more difficult to indulge in, because anyone in the EU may simply subscribe to them, rather than shop for (or even cook) their own "ethics".

3 Ethics Bluewashing

In environmental ethics, *greenwashing* (Delmas and Burbano 2011) is the malpractice of a private or public actor seeking to appear greener, more sustainable, or ecologically friendlier than it actually is. By "ethics bluewashing", I mean to refer to the digital version of greenwashing. As there is no specific colour associated with ethically good practices in digital technologies, "blue" may serve to remind one that we are not talking about ecological sustainability but mere digital ethics cosmetics:[6]

Ethics bluewashing = $_{\text{def.}}$ the malpractice of making unsubstantiated or misleading claims about, or implementing superficial measures in favour of, the ethical values and benefits of digital processes, products, services, or other solutions in order to appear more digitally ethical than one is.

Ethics greenwashing and bluewashing are forms of misinformation, often achieved by spending a fraction of the resources that would be needed to tackle the ethical problems they pretend to address. They concentrate on mere marketing, advertising, or other public relations activities (e.g. sponsoring), including the setting up of advisory groups that may be powerless or insufficiently critical. Both malpractices are tempting because, in each case, the goals are many and all compatible:

(a) Distract the receiver of the message—usually the public, but any shareholders or stakeholders may be the target—from anything that is going wrong, could go better, or is not happening but should;
(b) Mask and leave unchanged any behaviour that ought to be improved;
(c) Achieve economic savings; and
(d) Gain some advantage, e.g. competitive or social, for example in terms of "good will".

However, contrary to what happens with greenwashing, bluewashing can more easily be combined with digital ethics shopping: a private or public actor shops for the principles that best fit its current practices, publicises them as widely as possible and then proceeds to bluewash its technological innovation without any real improvement, much lower costs, and some potential social benefits. These

[5] See also (Mazzini forthcoming).
[6] This is not to be confused with the term bluewashing "[…] used to criticize the corporate partnerships formed under the United Nations Global Compact initiative (some say this association with the UN helps to improve the corporations' reputations) and to disparage dubious sustainable water-use projects" (Schott 4 February 2010).

days, ethics bluewashing is especially tempting in the context of AI, where the ethical issues are many, the costs of doing the right thing may be high, and normative uncertainity or sometimes confusion are widespread.

The best strategy against bluewashing is the same already adopted against greenwashing: *transparency* and *education*. Public, accountable, and evidence-based transparency about good practices and ethical claims should be a priority on the side of the actors wishing to avoid the appearance of engaging in any bluewashing malpractice. Public and factual education, on the side of any target of bluewashing—not just the general public but also members of executive boards and advisory councils, for example—about whether and what effective ethical practices are actually implemented means that actors may be less likely to (be tempted to) distract public attention away from the ethical challenges they are facing.

As we recommended in Floridi et al. (2018), the development of metrics for the trustworthiness of AI products and services (and of digital solutions in general) would enable the user-driven benchmarking of all marketed offerings and facilitate the detection of mere bluewashing, improving public understanding, and engendering competitiveness around the development of safer, more socially and environmentally beneficial products and services. In the longer term, a system of certification for digital products and services could achieve what other similar solutions have achieved in environmental ethics: make bluewashing as visible and shameful as greenwashing.

4 Ethics Lobbying

Sometimes, private actors (are at least suspected to) try to use self-regulation about the ethics of AI in order to lobby against the introduction of legal norms, or in favour of their watering down or weakening their enforcement, or in order to provide an excuse for limited compliance. This specific malpractice affects many sectors, but it seems more likely in the digital one (Benkler 2019), where ethics may be exploited as if it were an alternative to legislation, and in the name of technological innovation and its positive impact on economic growth, a line of reasoning that cannot be easily supported in environmental or biomedical contexts. Here is a more general definition:

Digital ethics lobbying = $_{\text{def.}}$ the malpractice of exploiting digital ethics to delay, revise, replace, or avoid good and necessary legislation (or its enforcement) about the design, development, and deployment of digital processes, products, services, or other solutions.

One may argue that digital ethics lobbying is a poor strategy, likely to fail in the long run because it is at best short-sighted: sooner or later legislation tends to catch up. Whether this argument is convincing or not, digital ethics lobbying as a short-term tactic may still cause much damage, by delaying the introduction of necessary legislation, by helping manoeuvre around or by-pass more demanding interpretations of current legislation, thus making compliance easier but also misaligned with the spirit of the law, or by influencing law-makers to pass legislation that is more favourable to the lobbyist than would otherwise be expected. Furthermore, and very importantly, the malpractice, or the suspicion of it, risks undermining the value of any digital ethical self-regulation *tout court*.

This collateral damage is deeply regrettable because self-regulation is one of the main valuable tools available for policy-making. In itself, it cannot replace the law, but if properly implemented, it can be crucially complementary (Floridi 2018), when:

- Legislation is unavailable (for example, in experimentations about augmented reality products) or
- Legislation is available, but also in need of an ethical interpretation (for example, in terms of understanding a right to explanation in the GDPR) or
- Legislation is available, but also in need of some ethical counterbalancing:

- If it is better not to do something, even if it is not (yet) illegal to do it (for example, to automate entirely and fully some medical procedure without any human supervision) or
- If it is better to do something, even if it is not (yet) legally required (for example, to implement better labour market conditions in the Gig Economy).

The strategy against digital ethics lobbying is twofold. On the one hand, it must be counteracted by good legislation and effective enforcement. This is easier if the lobbying actor (private or public) is less influential on law-makers or whenever public opinion can exercise the right level of ethical pressure. On the other hand, digital ethics lobbying must be exposed whenever it occurs and be clearly distinguished from genuine forms of self-regulation. This may happen more credibly if the process is also in itself part of a self-regulatory code of conduct of a whole industrial sector, in our case the digital tech industry, which has a more general interest in maintaining a healthy context where genuine self-regulation is both socially welcome and efficacious and ethics lobbying is exposed as unacceptable.

5 Ethics Dumping

"Ethics dumping" is an expression coined in 2013 by the European Commission to describe the export of unethical research practices to countries where there are weaker (or laxer, or perhaps just different, in the case of digital ethics) legal and ethical frameworks and enforcing mechanisms. It applies to any kind of research—including research in computer science, data science, machine learning, robotics, and other kinds of AI—but it is most serious in health-related and biological contexts. Fortunately, biomedical and environmental ethics may be considered universal and global; there are international agreements and frameworks and international institutions monitoring their application or enforcement, so "ethics dumping" may be fought more effectively and coherently when research involves biomedical and ecological contexts. However, in digital contexts, the variety (or indeed the lack of) of legal regimes and ethical frameworks facilitates the export of (what are considered within the original context where the "dumper" operates) unethical (or even illegal) practices, and the import of the outcomes of such practices. In other words, the problem is twofold, about *research ethics* and *consumption ethics*. So, here is a definition:

Digital ethics dumping = $_{\text{def.}}$ the malpractice of (a) exporting research activities about digital processes, products, services, or other solutions, in other contexts or places (e.g. by European organisations outside the EU) in ways that would be ethically unacceptable in the context or place of origin and (b) importing the outcomes of such unethical research activities.

Both (a) and (b) are important. To offer a concrete, if distant, example, it is not unusual for countries to ban the cultivation of genetically modified organisms, but allow their import. This asymmetry of ethical (and legal) treatment between a practice (unethical and/or illegal research) and its output (ethically and legally acceptable consumption of the output of the unethical research) means that ethics dumping may affect digital ethics not only in terms of unethical export of research activities but also in terms of unethical import of the outcomes of such activities. For example, a company may export its research and then design, develop, and train algorithms, e.g. for face recognition, on local personal data in a non-EU country with a different or weaker ethical and legal framework for personal data protection, which would be unethical and illegal in the EU because of the GDPR. Once trained, the algorithms may then be imported to the EU and deployed without incurring any penalty or even be frowned upon. Whereas the first step (a) may be more easily blocked, at least in terms of research ethics (Nordling 2018); the second step (b), involving the consumption of unethical research results, is fuzzier, less visibly problematic, and hence more difficult to monitor and curtail.

Unfortunately, it is likely that, in the near future, the problem of digital ethics dumping will become increasingly serious, due to the profound impact of digital technologies on health and social care as well as defence, policing and security, the ease of their global portability, the complexity of the productions processes (some stages of which may involve ethical dumping), and the immense economic interests at play. For example, especially in AI, where the EU is a net importer of solutions from the USA and China, private and public actors risk not just exporting unethical practises but also (and independently) importing solutions that may have been developed in ways that would not have been ethically acceptable within the EU.

In this case too, the strategy is twofold. One must concentrate on research ethics *and* the ethics of consumption. If one wishes to be coherent, both need to receive equal attention.

In terms of research ethics, it is slightly easier to exercise control at the source, through the ethical management of public funding for research. In this, the EU is in a leading position. However, there remains the significant problem that much R&D about digital solutions is done by the private sector, where funding may be less constrained by geographical borders (a private actor can more easily relocate its R&D to an ethically less demanding place, a geographical variation of the ethics shopping seen in Section 2) and is not ethically scrutinised in the same way as publicly funded research.

In terms of consumption ethics, especially of digital products and services, much can be done both by the establishment of a system of certification for products and services that could inform procurement, as well as public and private use. As in the case of bluewashing, the reliable and ethically acceptable provenance of digital systems and solutions will have to play an increasing role in the following years if one wishes to avoid the hypocrisy of being careful about research ethics in digital contexts and yet relaxed about the unethical use of its outcomes.

6 Ethics Shirking

Ethicists are well acquainted with the old malpractice of applying double standards in moral evaluations. By applying a lenient and a strict approach, one can evaluate and treat agents (or their actions, or the consequences of their actions) differently than similar agents (actions or consequences), when in fact they should all be treated equally. Usually, a risk of double standards is based, even inadvertently, on bias, unfairness, or selfish interest. The risk I wish to highlight here belongs to the same family, but it has a different genesis. To highlight its specificity, I shall borrow the expression "ethics shirking" from the financial sector[7] and define it thus:

Ethics shirking = $_{\text{def.}}$ the malpractice of doing increasingly less "ethical work" (such as fulfilling duties, respecting rights, and honouring commitments) in a given context the lower the return of such ethical work in that context is mistakenly perceived to be.

Ethics shirking, like ethics dumping, has historical roots and often follows geopolitical outlines. Actors are more likely to engage in ethics dumping and shirking in contexts where disadvantage populations, weaker institutions, legal uncertainties, corrupted regimes, unfair power distributions, and other economic, legal, political, or social ills prevail. It is not unusual to map, correctly, both malpractices along the divide between Global North and Global South, or to see both as affecting above all Low- and Middle-Income Countries. The colonial past still exerts a disgraceful role. It is also important to recall that, in digital contexts, these malpractices can affect segments of a population within the Global North. The Gig Economy may be seen as a case of ethics shirking within developed countries. And the development of self-driving cars may be interpreted as an instance of research dumping in some states of the USA. In this case, the 1968 Vienna Convention on Road Traffic, which establishes international principles to govern traffic laws, requires that a driver is always fully in control and responsible for the behaviour of a vehicle in traffic. However, the USA is not a signatory country and the requirement does not apply, meaning state vehicle codes do not prohibit automated vehicles, and several states have enacted laws for automated vehicles. This is also why research on self-driving cars happens mostly in the USA—as well as the related incidents and human suffering.

The strategy against ethics shirking consists in tackling its origin, which is a lack of clear allocation of responsibility. Agents may be more tempted to shirk their ethical work in a given context the more they (think they) can relocate responsibilities elsewhere. This happens more likely and easily in "D contexts", where one's own responsibility may be perceived (mistakenly) to be lower because it is *distant, diminished*, *delegated*, or *distributed* (Floridi 2013). Thus, ethics shirking is an agency unethical cost of deresponsabilisation. It is this genesis that makes it a special case of the ethical problem of double standards. This is why more fairness and less bias are necessary—insofar as ethics shirking is a special case of the problem of double standards—but they are also insufficient to remove the incentive to engage in ethics shirking. To uproot such a malpractice, one also needs an ethics of distributed responsibility (Floridi 2016) that relocates responsibilities—and hence praise and blame, reward and punishment, and ultimately causal accountability and legal liability—where they rightly belong.

[7] https://www.nasdaq.com/investing/glossary/s/shirking I owe the suggestion to include "ethics shirking" as a signifícnt risk in digital ethics and to use the expression itself to capture it to (Cowls, Png, and Au unpublished).

7 Conclusion

I hope this short article may work as a map for those who wish to avoid or minimise some of the most obvious and significant ethical risks, when navigating from principles to practices in digital ethics. From a Socratic perspective, a malpractice is often the result of a misjudged solution or a mistaken opportunity. Understanding as early as possible that shortcuts, postponements, or quick fixes do not lead to better ethical solutions but to more serious problems, which become increasingly difficult to solve the later one deals with them, does not guarantee that the five malpractices analysed in this article will disappear, but it does mean that they will be reduced insofar as they are genuinely based on misunderstanding and misjudgements. Not knowing better is the source of a lot of evil.[8] So, the solution is often more and better information for all.

References

Algorithm Watch. 2019. "The AI Ethics Guidelines Global Inventory." *https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/*.

Benkler, Y. (2019). Don't let industry write the rules for AI. *Nature, 569*(161). https://doi.org/10.1038/d41586-019-01413-1.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and Engineering Ethics, 24*(2), 505–528.

Floridi, Luciano, and Josh Cowls. forthcoming. A unified framework of principles for AI in society.

Cowls, Josh, Marie-Thérèse Png, and Yung Au. unpublished. Some tentative foundations for "Global" algorithmic ethics..

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review, 54*(1), 64–87. https://doi.org/10.1525/cmr.2011.54.1.64.

European Commission. 2019. "Ethics Guidelines for Trustworthy AI." *https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai*.

Floridi, L. (2013). Distributed morality in an information society. *Science and Engineering Ethics, 19*(3), 727–743.

Floridi, L. (2015). *The ethics of information*. Oxford: Oxford University Press.

Floridi, L. (2016). Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160112.

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180081. https://doi.org/10.1098/rsta.2018.0081.

Floridi, Luciano. 2019. "Establishing the rules for building trustworthy AI." Nature - Machine Intelligence.

Floridi, Luciano, and Tim Lord Clement-Jones. 2019. "The five principles key to any ethical framework for AI." New Statesman https://tech.newstatesman.com/policy/ai-ethics-framework.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, Effy Vayena, and %J Minds Machines. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. 28 (4):689–707. https://doi.org/10.1007/s11023-018-9482-5.

Mazzini, Gabriele. forthcoming. A system of governance for artificial intelligence through the lens of emerging intersections between AI and EU law. in *Digital revolution - new challenges for law*, edited by a. De Franceschi, R. Schulze, M. Graziadei, O. Pollicino, F. Riente, S. Sica and P. Sirena. SSRN: https://ssrn.com/abstract=3369266.

Nordling, L. (2018). Europe's biggest research fund cracks down on "ethics dumping". *Nature, 559*(7712), 17.

Samuel, A. L. (1960). Some moral and technical consequences of automation—a refutation. *Science, 132*(3429), 741–742.
Schott, Ben. 2010. "Bluewashing." The New York Times.
Wiener, N. (1960). Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358.
Winfield, Alan. 2019. "An updated round up of ethical principles of robotics and AI." *http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html*.