11 TOWARDS WISE OBJECTS: THE VALUE OF KNOWING WHEN TO QUIT

Pim Haselager

Constructive ethics considers the ethical, legal and societal implications (ELSI) of artificial intelligence (AI) in order to elucidate what will become possible (and when), what is desirable or what should be avoided and what should be regulated, by whom, when and how. An early identification of the concerns of stakeholders may help to identify a technology's main risks. Liability concerns are crucial to a developing technology. Smart objects (SOs) will increasingly operate in (1) dynamic and unpredictable situations, (2) with a variety of agents with different cognitive and behavioural capacities, (3) as part of teams with unclear responsibility transitions, (4) with users that can experience an inaccurate sense of agency and that (5) intentionally or unintentionally, may not always work towards objectively desirable goals. Therefore, I will make a plea for the development of 'wise objects': SOs that know when to protect their users by switching themselves off.

Introduction: On the ethical, legal and societal implications of smart objects

Developments in AI and cognitive neuroscience (CNS) take place with an ever-increasing speed. Their enormous effects on daily life practices and society at large (from communication practices of individuals to ways of working and earning wages) necessitate an investigation of their ELSI. It is important that such ELSI analyses do not stand 'outside of' the science and technology but rather aim to be interactive. This implies that ELSI analyses should be the result of, and provide

input for, communication with scientists and designers and potential stakeholders about what is possible, desirable or avoidable. This way, constructive ethics is part of the research and design process, instead of merely providing ethical codes before, or evaluations after, research and development (R&D). Ethicists working in the domains of SOs would therefore, I suggest, do well to engage in a continuous cycle of listening to scientists and designers, analyse presuppositions or implications of ongoing R&D, inform practitioners and stakeholders of potential consequences, and ask them about possibilities for change or improvement. From this perspective, ethics is not about telling others what should (not) be done, nor to instruct researchers to 'be good'. Instead, it can clarify or raise issues that require reflection and stimulate discussion, the results of which, sometimes, can be integrated into research. This will lead not only to ethically more prudent technology but also to commercially more successful products. After all, applications that, through their design, forestall or avoid concerns of clients and stakeholders will be more acceptable to users.

For this reason, it is important to ask several basic ELSI questions as early as possible in the research and design process. The first question is simply: What is possible? This question is a complicated one because often is not easy to separate the hope from the hype, and because time frames are regularly not clearly specified. Regarding the first, issue, companies like Gartner (https://www.gartner.com/en/research/methodologies/gartner-hype-cycle) produce regular analyses of developments in the field of AI and neuroscience, that indicate the estimated location of developing applications in a so-called 'Hype cycle'. They suggest for a variety of developing technologies whether they have been just triggered, suffer from inflated expectations, go through a period of disillusionment because of not fulfilling these expectations or finally reaching a level of useful productivity. For instance, in 2016 the Internet of Things (IoT) was located at the peak of the inflated expectations, because it was estimated to be still far removed from becoming genuinely practically applicable, while being much discussed in popular media (https://blogs.gartner.com/smarterwithgartner/files/2016/11/Hype-Cycle-for-the-Internet-of-Things-2016_Infographic-01.png). Although, of course, there is room for argument or disagreement regarding the exact location of a particular technology (or even the exact meaning and extension of a particular label), the fact that much advertised or discussed technologies may contain a large amount of overpromise complicates a proper reflection on the ethical implications of such technologies. Just like the products they are about, ELSI analyses could be hyped too, in the sense that they exaggerate prematurely the risks associated with a certain type of application.

Moreover, it is important to specify the temporal path that is being considered in an ELSI analysis. Many discussions about the hopes, concerns or risks of a technology fail to indicate a clear timeframe they consider in discussing this technology and its associated risk assessment. Without attempting to be overly

precise, it generally is useful to distinguish what is currently possible (actually existing applications) from what will be possible in the near future (currently discussed in research papers, with the potential to develop applications in the next 5 years). These, in turn, have to be separated from an estimate of what would be possible in the long run (say the next twenty-five years or so), which has to be distinguished again from fantasizing about what 'ultimately' might become possible (science fiction).

A second set of questions concerns the assessments of stakeholders regarding what is desirable ('dreams'), what should be avoided ('nightmares') and how the technologies involved should be stimulated or restricted, via funding, regulations or laws. Gupta, Fisher and Frewer (2011) indicate that perceived risk is one the most frequently investigated socio-psychological determinants of public acceptance of a technology, much more than trust or perceived benefit. This implies that a proper risk analysis of a developing or to-be-developed technology will be crucial, right from the start. I will come back to this issue in the next section.

A third set of questions revolves around the identification and consultation of stakeholders in technology, which can be a complex and time-consuming task. Just to give a small list of candidates to be considered in general: end users, researchers, scientific experts or organizations (e.g. in relation to self-regulation via ethical codes), patient groups, caregivers, policymakers (governments), legal institutions, non-governmental organizations, companies developing the technology, companies using the technology, insurance companies, the general public and undoubtedly many other candidates. It is for this reason that ELSI analyses should start early in the research and design process. Even though this runs the risk of ethically assessing technology that is not there yet, societal debates about what to pursue or avoid, about what to regulate and why, take time too. Waiting for the technology to be out on the market before societal debates about ELSI take place runs the risk of being too late to effectively diminish the potentially negative effects of the technology. One example that comes to mind the loss of privacy on the internet, which at least in part can be seen as a consequence of not discussing early enough, for example, the effects of tracking cookies on websites. Although serious attempts are now being made to restore some of the privacy lost, the difficulties involved in this endeavour clearly illustrate the danger of engaging in ELSI analyses too late.

On the creative use of SOs by different agents in dynamic environments

As indicated above, perceived risks are an important element in the evaluation of a developing technology. An important factor in the risk assessment of SOs is that they will operate in a dynamic, unpredictable environment. As the autonomy

and intelligence of SOs increase, these products will be capable of dealing with increasingly more complex and demanding tasks. Using AI 'in the wild' (Hutchins, 1995) implies that the SOs will operate in messy and quickly changing surroundings. In general, it is hard if not impossible to foresee exactly how such SOs will behave in dynamic environments. In the field of robotics, for example, prototypes get tested extensively in lab renditions of real environments. But so far, it is, at least to my knowledge and experience, rarely the case that such robots are tested in real-life circumstances under unrestricted or uncontrolled conditions for such prolonged periods of time that a complete, full-scale assessment is possible. Simply put, the world is just too big, complex and dynamic to exhaustively test relatively intelligent and autonomous systems.

Second, the label 'Human-Robot Interaction' (HRI) does not do justice to the ecologies of SOs, robots and various other types of intelligent agents that will interact in real-life settings. Just taking an elderly care institution as an example, studies usually focus on how the elderly person will interact with the robot, and even, in some cases, the caregiver as well. But in real life it is to be expected that robots will have to interact with a variety of other agents, for example, elderly, adults, patients, children, animals and other robots, with each having their own individual and behavioural (in)capacities, interactive styles and preferences. Here, too, tests will run short by far of the complexities to be expected in real-life robot interaction with a variety of other agents.

Third, speaking of a 'caregiver' may be an oversimplification, given the various parties involved in care contexts that each, to some extent, may be responsible for and/or determining parts of the robot's behaviours. In many situations care is not the responsibility of an individual, but of a team. For instance, in a caregiving context, the institution may be the owner of the SOs (say, a robot), local technicians may help to setup the robot for the specific environment it should operate in, professional caregivers may be instructors as well as collaborators of the robots, while the elderly are end users who instruct and experience the effects of the robot's actions. It will not be immediately clear, or even difficult after the fact, which agents carry which parts of the responsibility for the effects of an SO's action.

Fourth, it need not always be easily determinable to which extent users of smart technology will be aware of their actual use of it, not even for those users themselves. Increasingly, smart technology moves towards becoming symbiotic systems, becoming combined with or encompassing their users, rather than being used by them. In the contexts of brain-computer interfacing (Haselager, 2013; Krol, Haselager & Zander, 2019), as well as shared AI-user control (Vilaza, Haselager, Campos & Vuurpijl, 2014; Abbink *et al.*, 2018), situations may occur where users are unaware that they are acting or are incorrect in their experience of a sense of agency. That is, users may misinterpret their contributions to the performance of an action, either underestimating or exaggerating their contributions.

Finally, it may be incorrect to assume that under ordinary conditions, the user will always have beneficial intentions regarding the use of the SO. In some cases, the potential negative consequences of the use may be unintentional. A classic example is that of a cat sitting on a Roomba, attacking a pit bull (https://www.youtube.com/watch?v=vf9wHkkNGUU). The function of this SO, that is, to clean the floor, was made more difficult (by the cat's weight) and its actions were to some extent appropriated by the cat for different purposes, or at least leading to unplanned behavioural improvisations, such as attacking the dog. In addition to such accidental improvised use, the behaviour of bystanders or users could even go intentionally against the SO's functions, or the SO itself. This has been observed, for example, in the mal-education of chatbot 'Tay' (Vincent, 2016) that deliberately was taught to make racist and sexist statements, or in cases of robot bullying (Salvini *et al.*, 2010), where humans attack and damage robots. In an increasingly smart ecology, it seems rather likely that unforeseen, unplanned or even intentional 'creative use' with potentially negative consequences will manifest itself.

These five aspects of 'creative use' of SOs may have legal consequences, for example, in assessing the liability for the consequences of an SO's actions. For instance, Asaro (2012, p. 171) has pointed out the complexities of applying product liability in the context of smart systems, in his case, robotics:

> Legal liability due to negligence in product liability cases depends on either *failures to warn*, or *failures to take proper care* in assessing the potential risks a product poses. … What constitutes proper care, and what risks might be foreseeable, or in principle unforeseeable, is a deep and vexing problem. This is due to the inherent complexity of anticipating potential future interactions, and the relative autonomy of a robotic product, once it is produced.

In tort law, strict liability concerns the imposition of liability on a party without a finding of fault (negligence or intention). The law imputes strict liability to situations it considers to be inherently dangerous: defective products, dangerous tools. Here the question arises how to evaluate SOs working within a dynamic and unpredictable environment (including e.g. cats, pit bulls and potentially agents with bad intentions as in the examples above). Should or could such an SO be considered as potentially 'defective' or 'dangerous', at least under some circumstances?

A different but related issue comes to the fore when one considers Article 6 of the EU Directive (85/374/EEC; see also Dodds-Smith, 2017), which states,

> A product is defective when it does not provide the safety which a person is entitled to expect, taking all circumstances into account, including:
>
> (a) the presentation of the product;

(b) the use to which it could reasonably be expected that the product would be put;
(c) the time when the product was put into circulation.

Just how much safety is a person entitled to expect regarding smart products, and what uses can SOs be reasonably expected to be put to? It does not seem that a framework for answering these questions in relation to the development of SOs exists. Yet it is unlikely that standard ways of addressing these questions within the contexts of traditional, non-smart objects will easily be translatable to this new domain. Perhaps even more important than directly answering these questions is the question whether the field of SOs is *aware* of such liability issues. To what extent are such considerations part of the research and design processes?

Midas's touch and cars with parachutes

The consequences of being unaware (or neglecting) issues surrounding responsibility for SOs and potential product liability can be vital. Awareness of these consequences is important not just for the development of safe products but also for the societal acceptance of these products, or even for the legitimate introduction of such products on the market. Hence, I suggest that objects should be more than 'smart'. What we need are 'wise objects' in the sense of carrying within their design a control mode related to responsibility and liability concerns. Smart technology may be said to be in danger of having the so-called Midas touch, in that it, once operative, may affect everything *continuously*. As Turkle (2008, p. 2) indicates, 'We are tethered to our "always-on/always-on-us" communication devices … always ready-to-mind and hand.' This raises questions, once the technology is on, who can turn it off again, when and how? Should the system itself be equipped with some kind of 'emergency brake'? What would be the most graceful way of building emergency brakes into the system?

This presents a challenge that can be perceived as paradoxical, especially by engineers and designers. In the first stages of developing a product, all the attention and work is generally directed at creating and improving the SOs main function (whatever it may be). Yet, from an ELSI perspective, it is precisely during those early stages that one should think about its complete opposite, namely preventing it from continuing to do whatever it is doing by turning it off. Just as a simple example, when designing the motor of a car, it is logical to focus on improving the motor, because driving fast(er) is obviously the car's main function. Yet, from the perspective of stakeholders (ranging from drivers to passengers and other traffic participants to insurance companies and legislative institutions), the capacity to stop might be considered to be even more important than the capacity to drive. The paradox implies that as soon one has thought of a rudimentary implementation

of the basic function of the designed object, one needs to start addressing its opposite: a functionally easy to use, 100 per cent reliable, off-switch. Ignoring this issue might lead to suboptimal additions to the design later, for example, a car with only a parachute as its main brake.

In several cases, of course, risks of a product or technology are apparent enough to be taken into account from the start. Early on in the design process of cars, the need to be able to stop at any time was too obvious to be ignored, preventing later ad hoc patches like parachutes. However, it is far from clear that in the case of SOs operating in increasingly real-life environments, in an increasingly autonomous way, such risks are as apparent as in the case of cars. Hence, it is vital that serious attention (and some creativity) is applied to considering the types of risks SOs might bring, and the types of emergency brakes that are required to diminish such risks. Indeed, just as an illustration, within the EU, debates have started about the requirement for 'kill switches' on robots (Kottasova, 2017). Part of the challenge of designing wise objects lies in the development and acceptance of a framework for addressing risk, and brake-related issues. How can a technology be stopped as immediate and as fail-proof as possible? Who should be able to decide when to turn it off? Who would be held accountable for turning it off (or not)? Such questions are not to be seen as (later) additions to, but as an integral part of, the design process to be addressed and solved early on.

Wise objects

One option that presents itself in the context of SOs is that they are designed such that instead of (or preferably in addition to) having reliable ways to be stopped, SOs could be designed to turn off themselves. Just like a wise person knows when to shut up, a wise object should know when to stop. Simple examples exist of robots that can or should say 'No' to their users (Briggs & Scheutz, 2015; Förster, Saunders & Nehaniv, 2018; Peeters & Haselager, 2019). To what extent would it be possible to design SOs that can say 'No' to themselves? This might be important when human intervention would be too late to be meaningful, as the damage may already have been done, or too late to be effective, as when a rogue system spinning around at high speed would make an emergency brake unreachable (Arnold & Scheutz, 2018). It could be even more relevant to have self-terminating SOs in cases where their users require protection in relation to, for instance, consequences that the user is not, or insufficiently, aware of. This plea for wise objects is part and parcel of the broader perspective of beneficial AI (e.g. Russell, 2017).

Is it possible, for instance, to design SOs that collect data in order to provide better services to block data transfer for further processing (or even delete the data) when local analysis (within the SO) indicates a privacy risk? Ideally, wise objects should function as virtuous guardians of a user's privacy instead of the

data-sucking vampires that they currently often are. Obviously, there are many challenges here. What criteria would have to be used, what threshold settings would function reasonably enough often enough? Which application domains require such wise objects most urgently? SOs for young children, for example, smart toys, might provide an interesting domain for further study.

Wise objects are SOs that can be trusted. As van den Brule *et al.* (2014, 2016) indicate, trust can be defined as the willingness of one agent (the 'trustor') to be vulnerable to the actions of another agent (the 'trustee'). The trustor depends on the trustee to reach its goals, but there is risk to the trustor if the trustee's actions fail or betray. In current SOs the risks of failure and/or betrayal can be unacceptably high. The most pressing challenges in design therefore don't involve the development of increasingly smarter objects, but rather of objects that are wise.

Conclusion

A consideration of the typical ELSI questions regarding what is possible when, and what possibilities should be strived for or avoided, quickly leads designers to concern themselves with the potential risks of their products. In the case of SOs significant risks derive from five factors inherent in their aimed for functionality. The more smart and autonomous objects will be, the more they will function for prolonged periods of time in dynamic and unpredictable circumstances, dealing with increasingly complex tasks. In those situations, it is likely that the SOs will have to interact with not just one user but with changing ecologies of users, SOs and various other types of agents with different cognitive and behavioural (in)capacities. Because the design, setting up, training, collaboration and use of SOs often involves many different agents, attributing (aspects of) legal responsibility may become quite complex. Sometimes, human users may not be fully aware of their (lack of) agency when using smart technology. In other cases, the use of SOs can be 'creative', in the sense of unintentionally or intentionally working towards negative consequences in not easily foreseeable or humanly preventable ways. I have therefore suggested that it is important to focus, from the start of the design process, on working towards wise objects: SOs that know when to quit.

Bibliography

Abbink, D. A., Carlson, T., Mulder, M., Winter, J., Aminravan, F., Gibo, T., *et al.* (2018). A topology of shared control systems: Finding common ground in diversity. *IEEE Transactions on Human-Machine Systems*, *48*(5), 509–525. doi: 10.1109/THMS.2018.2791570.

Arnold, T., & 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*, 59–69.

Asaro, P. M. (2012). A body to kick, but still no soul to damn: Legal perspectives on robotics. In P. Lin, K. Abney & G. A. Bekey (Eds), *Robot ethics: The ethical and social implications of robotics* (pp. 169–186). Cambridge, MA: MIT Press.

Briggs, G., & Scheutz, M. (2015). 'Sorry, I can't do that': Developing mechanisms to appropriately reject directives in human-robot interactions. In *Proceedings of the 2015 AAAI Fall Symposium on AI and HRI*.

Brule, R. van den, Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., & Haselager, P. (2014). Do robot performance and behavioral style affect human trust? *International Journal of Social Robotics*, *6*, 519–531.

Brule, R. van den, Bijlstra, G., Dotsch, R., Haselager, P., & Wigboldus, D. H. J. (2016). Warning signals for poor performance improve human-robot interaction. *Journal of Human-Robot Interaction*, *5*(2), 69–89.

Dodds-Smith, I. (2017, 21 May). Recent developments in European product liability. *Arnold&Porter*. Retrieved 22 October 2020 from https://www.arnoldporter.com/en/perspectives/publications/2017/05/recent-developments-in-european-product-liability.

EU Council directive 25-07-1985, (85/374/EEC). Retrieved from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A31985L0374.

Förster, F., Saunders, J., & Nehaniv, C. L. (2018). Robots that say 'no': Affective symbol grounding and the case of intent interpretations. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(3), 530–544.

Gupta, N., Fischer, A. R. H., & Frewer, L. J. (2011). Socio-psychological determinants of public acceptance of technologies: A review. *Public Understanding of Science*, *21*(7), 782–795.

Haselager, W. F. G. (2013). Did I do that? Brain–computer interfacing and the sense of agency. *Minds & Machines*, *23*(3), 405–418.

Haselager, P., Mecacci, G., & Wolkenstein, A. (forthcoming). Can BCIs enlighten the concept of agency? A plea for an experimental philosophy of neurotechnology. In O. Friedrich, A. Wolkenstein, C. Bublitz, R. J. Jox & E. Racine (Eds), *(Clinical) neuroethics meets artificial intelligence: Philosophical, ethical, legal and social implications*. Cham: Springer.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Kottasova, I. (2017, 12 January). Europe calls for mandatory 'kill switches' on robots. *CNN*. Retrieved from https://money.cnn.com/2017/01/12/technology/robot-law-killer-switch-taxes/index.html.

Krol, L. R., Haselager, W. F. G., & Zander, T. O. (2020). Cognitive and affective probing: A tutorial and review of active learning for neuroadaptive technology. *Journal of Neural Engineering*, *17*(1). https://doi.org/10.1088/1741-2552/ab5bb5.

Peeters, A., & Haselager, W. F. G. (2019). Designing virtuous sex robots. *International Journal of Social Robotics*, *13*, 55–66.

Russell, S. (2017). Provably beneficial artificial intelligence. In *The next step: Exponential life* (pp. 178–192). BBVA OpenMind. Retrieved from https://www.bbvaopenmind.com/en/books/the-next-step-exponential-life/.

Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B., *et al.* (2010). How safe are service robots in urban environments? Bullying a robot. In *19th International Symposium in Robot and Human Interactive Communication* (pp. 1–7). Viareggio, Italy: IEEE.

Turkle, S. (2008). Always-on/always-on-you: The tethered self. In James E. Katz (Ed.), *Handbook of mobile communication studies* (pp. 121–137). Cambridge, MA: MIT Press

Vilaza, G. N., Haselager, W. F. G., Campos, A. M. C., & Vuurpijl, L. (2014). Using games to investigate sense of agency and attribution of responsibility. In *Proceedings of the 8th*

Brazilian Games and Digital Entertainment Symposium (SBGames), Porto Alegre (2014, Nov 12–14) (pp. 393–399). https://www.sbgames.org/sbgames2014/papers/culture/full/Cult_Full_Using%20games%20to%20investigate.pdf.

Vincent, J. (2016, 24 May). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. Retrieved 22 October 2020 from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.