Ethical Aspects and Challenges for Interactive Task Learning

Matthias Scheutz

4/17/18

1 Introduction

New technologies, such as being able to teach machines how to perform tasks through interaction instead of having to program them, are always exciting and have the potential to significantly benefit humanity. At the same time, every transformative technology raises important ethical questions, if adopted, about the risks involved and potential impact on human societies. Machine learning (ML) is such a recent, potentially transformative technology that poses many important ethical challenges for designers of learning algorithms: How can we ensure that the algorithm will learn what it is supposed to learn and only that? How can we be certain that the machine will perform correctly after it has learned something new? And how can we guarantee that the machine will not behave in unethical ways after learning?

These challenges arise with all forms of machine learning, and thus with *interactive task learning* (ITL) as well. The difference with ITL is that, in addition to ethical questions pertaining to the machine's newly acquired knowledge and how it will be used by the machine, there are questions regarding the learning interaction and how the interaction might affect humans subsequently. Moreover, there are questions about acquisition and use of the "normative surroundings" of the task, i.e., all the ethical principles and implications considered to be part of the task that might not be explicitly instructed, but have to be followed during task execution.

Interactive task learning, especially when it involves task-based natural language interactions, has many important advantages for humans and machines alike. For one, natural language instruction is a very intuitive modality for humans, since humans are used to teaching each other through natural language dialogues. Moreover, it requires less training and preparation on the human side (compared to instructing tasks, e.g., through some form of programming language), and it allows human instructors to check quickly whether the learner has taken in the lesson. For machines, ITL is advantageous because ideally it allows them to acquire high-level task descriptions very quickly, instead of having to construct them over long stretches of bottom-up abstractions from low-level data. Moreover, being able to ask the human instructor for help (e.g., explanations or alternatives for action) does not exist in a data-driven approach; if the answer is not in the data, no statistical process in the world will be able to extract it. However, there are potential downsides to ITL as well, in particular for humans. Humans would have to engage with an autonomous machine over possibly extended periods of time in a way that would lead to successful knowledge acquisition for the machine. Some effects would be limited to the teaching interaction itself, whereas others could extend far beyond the teaching context.

In the following, we examine three classes of ethical aspects as they arise in ITL:

1. What is being taught and what are the associated risks?
2. What are the dynamics of human-machine instruction?
3. What effects will ITL have on human instructors and society?

2 The Risks of Machine Learning

Whenever a machine is allowed to acquire new knowledge by way of employing its own learning algorithms (as opposed to having knowledge implanted by human engineers, which has its own challenges), there is always the risk that its learning process will not work as intended by the designers of the learning algorithm. ITL learning algorithms are no exception. The learning algorithm might not acquire the right kind of information, only acquire incomplete knowledge, or acquire more knowledge than intended. For example, consider the case of an ML algorithm that is supposed to learn how to detect human faces in pictures. This algorithm might fail to detect faces per se, but learn to detect humans instead (an instance of the first case); it may learn to detect faces but miss some faces under particular lighting conditions (an instance of the second case); or it may learn to detect faces as well as additional factors about humans such as their sexual orientation (an instance of the third case) (Wang and Kosinski 2017). Clearly, the result of unintended effects of ML can have ethical implications (e.g., when unintended and unexpected information is exposed that can be misused such as likely medical conditions or when wrongly classified information is used for making decisions that impact human lives such as denying credit loan applications).

With ITL, several additional aspects come into play that are typically not an issue with statistical machine learning from data, for ITL involves direct personal interactions between human instructors and machines, different from impersonal ML from data sets. For instance, if a data set does not contain information that the machine is supposed to learn, the data-driven ML algorithm is not at fault if the machine fails to extract it. In the case of ITL, however, it is not clear who is to blame when critical aspects of a task are not picked up by the machine: was the learning algorithm at fault because it failed to encode or infer the information, or was the human to blame because the information was not properly taught to the machine? Determining this will be difficult for the same reasons that ITL is challenging: How much knowledge can the human instructor assume that the machine has? How much language does the machine understand? How detailed do instructions have to be, and how much can the machine infer necessary aspects on its own? How can the instructor determine that the machine has fully acquired the task? Would it have to demonstrate it to the human or would repeating it back, maybe in its own words, do the trick? While it is likely that machines will get blamed (at least initially while ITL technology is still developing), the very possibility that a human instructor will get blamed (or might blame herself) if a machine fails to learn a task after instruction raises ethical questions about the nature and expectations of such interactions and the subsequent effects on the involved humans.

Based on empirical work in human-robot interaction, it seems reasonable to assume that cues of human-likeness (such as natural language understanding) will cause people to import a vast array of assumptions about the machines' capabilities (such as background knowledge and level of understanding) that may not be warranted. In fact, it is likely that despite sufficient perceptual capabilities and being able to understand enough natural language to be able to learn new tasks, machines capable of ITL will not be human-like in many other ways. Incorrect impressions formed from limited exposures during teaching sessions might lead humans to omit important aspects of tasks (e.g., relating to property ownership, personal liability, human rights, and others) which in turn could lead to unintended consequences during task performance (e.g., the machines might not pay attention to whether the objects they use belong to their owner or whether they impact the freedom of other agents in performing the tasks). Moreover, humans are known to be "sloppy" when providing instructions. They use imprecise terminology, leave gaps in descriptions, refer to wrong objects, and make other errors, none of which is typically not a problem for a human learner who can “think along” and automatically correct such errors (Thomaz et al, this volume).[1] This raises the question: To what extent will and can we expect machines to read between the lines, and who is to blame when machines fail to derive the correct human intention?

Any sort of misunderstanding between instructor and machine could result in the machine learning the wrong task, learning it incompletely, or not learning it at all. There is also, of course, the possibility that the machine learns the task in a way that the human did not intend. Consider the example of a search and rescue robot that is supposed to learn how to find wounded people after a natural disaster. The aim is for the robot to enter collapsed buildings and make its way through the rubble to find any humans trapped inside the structure. Now suppose the robot takes the goal to search for wounded people too literally, so much so that when it finds a non-injured person, it determines that it cannot report its discovery because the person is not wounded. Hence, to be able to report the human, the robot decides, on the spot, to inflict an injury on the person to enable reporting the person as wounded and collect the reward from its policy-based decision-making system or from its partial satisfaction planner. While this problem almost seems comically bizarre, it is actually hard to prevent such cases without explicitly providing constraints to the system about which actions a machine is not allowed to perform to achieve its task goals. But what would serve as the source of this type of knowledge? Who would be in charge of providing it, and who would be responsible for ensuring that the robot was able to apply it correctly?

This raises the question about the extent to which ethical aspects related to a task (e.g., task-based obligations and prohibitions, social norms to be followed while learning the task, ethical principles to be applied while executing it, etc.) need or ought to be taught as well because they cannot be assumed to be known to the machine. For example, it may seem banal to us that when instructions are given to a machine to build a fence, that the machine must obtain the ingredients legally and must not simply take them from the surrounding areas (e.g., dismantling the neighbor’s fence). Another example would be that when a Thanksgiving turkey dinner is to be prepared, the turkey ought to be dead already. Such constraints are obvious to us, hence they are typically not included in task instructions because we assume that our instructees know them and we rely on their ability to apply what a “reasonable person” would do. This legal term packs in a lot of common sense knowledge and reasoning that we, however, cannot assume machines will automatically possess. Hence, in addition to determining what legal and moral aspects of a task the instructor needs to explicitly teach, there might also be social normative aspects that the system would have to know to safely operate in human environments (e.g., to not quickly move around in crowded human environment, not moving towards people while holding knives pointing at them, etc.).

3 The Dynamics of Human-Machine Instruction

The fundamental difference between ITL and other task learning algorithms (e.g., learning tasks from instructional videos) is the real-time personal interaction with a human instructor. The human element imposes constraints and restrictions on the types of interactions machines may conduct with humans to be respectful of human normative expectations (e.g., politeness), but also

[1] For example, in a search and rescue task (performed in a past experiment in our lab) where a human director had to instruct a human searcher to find yellow blocks, the erroneous instruction to look for yellow “books” was not a deterrent for the searcher.

human cognitive abilities and limitations (e.g., memory constraints, limited focus of attention, etc.) – a problem that does not have to be addressed and is thus not part of data-driven learning algorithms. For example, exhibiting appropriate demeanor will be critical for machines to ensure humans will not be offended and thus unwilling to continue an interaction (e.g., for a machine to repeatedly drop comments like “easy enough” and “no problem” might be construed by the human as downplaying what might otherwise have been a difficult task for humans to acquire, e.g., learning how to play the “Flight of the Bumblebee”). This includes respecting social norms such as politeness norms that guide interactions among humans (e.g., if a robot needs a screwdriver at a certain point in the task and sees that the human instructor is holding one, it should not attempt to just take it without asking).

In addition to using the appropriate tone and attitude while interacting with humans, being respectful of human expectations and constraints is critical. Especially with early ITL systems, it is likely that they will not be able to meet human expectations (about natural language understanding, speed of actions, timing of interactions, etc.). For they will simply not be advanced enough in their interaction capabilities, background and common sense knowledge, and natural language understanding to learn in ways that humans would assume when instructing other humans (as discussed above, such expectations come naturally to humans when machines are appear to be human-like). As a result, teaching interactions will quickly devolve into unnatural and tedious exchanges for humans. It will be important to ensure that humans have the right *mental model* of their machine’s capabilities and that machines do everything they can to make their interactions less frustrating for humans. This will require machines to be aware of human emotional states and the effects different interaction patterns might have on human emotions (e.g., a robot repeatedly asking a human to rephrase an instruction because it could not understand it or because it was not precise enough will quickly frustrate the human).

Once ITL and machine capabilities have sufficiently advanced, being aware of human limitations will also become a critical component in order to preserve *human dignity*. For example, ignoring human cognitive limitations such as attention span, ability to remain focused and concentrate, speed of natural language processing and information intake, etc. can lead to dysfunctional interactions and overhead that does not serve either interactant well (e.g., a robot anticipating a human instruction after only a few words and start to execute it proactively might confuse people in the simplest case, but could even cause anxiety on the human side as the robot’s actions are not legible to the human).

Additional challenges arise with learning systems that might be able to alter or improve behaviors as they are being instructed (e.g., because their physical constraints or capabilities allow for alternative better ways of completing actions, possibly in a manner that would be impossible for humans) – how would those optimizations be received by humans and would humans be able to judge whether the system has been able to understand the task? E.g., imagine a robot that after having been instructed to follow the steps in a human instruction manual to assemble a drawer chest detects various shortcuts and alternative ways of grasping and assembling parts that it commences at a much faster speed than humanly possible, so that up until shortly before the job is done, the robot’s steps do not seem to make much sense to the human. The robot could also, by way of its physical capabilities, perform multiple steps in parallel (e.g., if it has several grippers it can use independently) or determine alternative ways of connecting parts that improve the stability of the chest (e.g., based on its own physics models or mental simulations). Such unexpected super-human performance might not only make the social interaction element uncomfortable and the teacher’s job of ensuring proper learning and performance harder, but may also leave psychological marks on the human that persist beyond the interaction, which we will discuss next.

4 The Extended Effects of ITL on Human Instructors and Society

Another important ethical aspect of new technologies is their longer-term impact on humans and human society. While ITL shares some of the same long-term questions with other machine learning approaches – how to ensure that machines will learn knowledge that they can put to good use, that they will serve humans well and not become deviant – it also has unique long-term ethical aspects that directly relate to human nature and need to be pointed out. Given that humans will interact with machines at the very least as teachers during the interactive learning process, it is important to ask whether this interaction could have potentially negative effects on humans beyond the teaching interaction. For example, will humans feel (possibly unnecessarily) responsible when machines did not manage to acquire a task properly? I.e., will humans blame themselves instead of the machine because they really cared about the machine's success? And will such caring when the machine after repeated learning interactions succeeds at its tasks prompt feelings of pride for the machine on the human side and possibly lead to the establishment of unidirectional emotional bonds (e.g., Scheutz 2012) because the human feels a personal connection with the machine?

Conversely, will the human be shocked, put off, or worried when observing machines with "super-human" task learning or task performance capabilities? A machine might determine, for example, that it does not have to stick with the performance limitations imposed by how the human taught it a task or how the human has to perform the task due to human sensorimotor constraints (e.g., a human might have to use a measuring tape in order to determine the length of a piece of wood that needs to be cut while a robot could immediately cut it, using its visual system to measure the correct length). Or consider machines that can consult cloud-based databases while interacting with human teachers, acquiring all necessary background knowledge quickly on the fly before a human instruction has even finished; or machines that may covertly exchange messages with other learning machines while they are being instructed to quickly learn news skills from those machines. Just take a robot that does not know how to use a drill which, as the human starts to explain how to operate drills, quickly assures the human that it just picked it up from consulting other robots in its cohort – there is evidence that humans find such covert communication disconcerting and eerie (Williams et al. 2015).

In general, we have to anticipate massive effects of machines that can rapidly learn new tasks from interactive instructions at the societal level. If performed with the right task representations and paired with knowledge-sharing, ITL could form the basis of massively parallel learning where teaching one machine means that all (connected) machines will know the task (Scheutz 2014). How such massive learning will affect labor markets and our economy is anybody's guess. But it seems reasonable to assume that the first-hand experience of such super-human performance by machines, the awe, but also the jealousy and inferiority we may feel when the machine rapidly perfects a skill, can have deep ramifications for how we as humans experience ourselves. In fact, it may lead to what the philosopher Günther Anders called the "Promethean Shame", the feeling of inadequacy resulting from watching our own technological products surpass us in their abilities and perfection, in particular, the realization that our capacity to think is inferior to that of our own machines (Anders 1956/1979).

5 Discussion

As with all new technologies, it is important to weigh the advantages and disadvantages of ITL, and to carefully consider the trade-offs and risks involved in allowing, or even requiring humans to teach machines. ITL certainly shares the worries expressed related to machine learning in the context of the larger discussion about the utility and dangers of artificial intelligence (AI) – how to guarantee that learning machines will be safe for humanity and how to ensure that they can be turned off if they evolved in a dangerous direction and everything else fails? These topics, currently discussed under the moniker "Big Red Button" (as a means to shut off deviant AI) apply to ITL in the same way they apply to other learning methods. But different from variants of reinforcement learning where machines have to be incentivized to let them be shut off (Orseau and Armstrong 2016), ITL allows for the explicit instruction of ethical principles in conjunction with tasks, an opportunity that if paired with the right computational architecture will make ITL a more desirable learning method for ensuring ethical behavior. Explicit instruction will also reduce the risk associated with placing all bets on the machine's ability to pick up normative principles from pure observation of human behavior, which may not be practical or even possible in the case of ITL (Arnold and Scheutz 2018).

Of course, instructing ethical principles along with tasks is putting the burden on ensuring ethical behavior on the human instructor, which then raises the question of who should be allowed to instruct machines? What if the instructor is not interested in providing ethical guidance, or simply does not have the knowledge to do so explicitly? Or even worse, what if the instructor has a malicious agenda, trying to instruct the machine how to build and place bombs? How would the machine know that that kind of task is off limits in most except for a few very specialized national defense contexts?

There is a tacit assumption underwriting the very idea of ITL that both teacher and learner will be benevolent, i.e., the teacher will not instruct inappropriate tasks and the learner only has the best human interests in mind (or at least not malicious intent). However, such assumptions may not be always warranted despite our best efforts: from teachers that may be unaware of task and/or environmental conditions that could make instructions ethically problematic, to instructions that are contradictory (where it is unclear how to resolve the conflict), to teachers who have ulterior motives to teach tasks incorrectly, to systems that have been compromised (e.g., by hackers) and try to coerce the human instructor into teaching them tasks they are not supposed to learn. Clearly, ITL cannot be considered in isolation from mechanisms in the computational architecture that prevent unethical behavior – allowing machines to blindly follow human instructions is a recipe for disaster.

In addition to all these challenges with ensuring the ethical behavior of instructible machines, ITL poses additional challenges due to the intrinsic involvement of human instructors that intersect closely related discussions in the ethics of human-robot interaction (HRI). Aside from the potential detrimental effects of ITL on the human psyche that have been anticipated by philosophers of technology for decades, there are questions about ownership, responsibility, and allegiance posed by ITL that have to be addressed: who should be allowed to teach a robot and what ought to be the limits of instruction? How is the robot supposed to handle "competing interests" (Arnold and Scheutz 2017) in social groups such as a family where multiple members might want to teach the robot different tasks? Whose orders should it follow? Who should be in charge for controlling what the robot is or is not allowed to learn and use? And who will assume responsibility for the robot's actions? There are currently no good answers for any of the above questions; for one, because the research communities in AI, robotics and HRI are still very much focused on understanding and addressing the fundamental technical challenges raised by ITL. Yet, what the above discussion hopefully demonstrated is that technical work on ITL cannot proceed in isolation from the ethical challenges raised by machines that can interactively learn new tasks.

6 Conclusion

As new research thrusts are emerging to advance the ability of machines to interactively learn from human instructors, it is imperative to keep the overarching ethical aspects pertaining to the ITL learning algorithms, the learning interaction between human teacher and machine learner, and the longer-term effects of the interaction on the human in mind to be able to deploy machines with ITL capabilities to the benefit of human societies. Far different from data-driven machine learning, which usually cannot get any normative context information out of training data, simply because that information is not contained in the data set, ITL offers the unique opportunity for explicit instructions of the "normative surroundings" of tasks: rules and regulations about task-relevant entities, social and moral norms associated with performing the task as well as other ethical principles involved in learning and performing the task. Machines instructed by ITL thus have the advantage to be able to co-learn their task with when, where, and how these tasks are appropriately performed. This, however, puts part of the onus on the human instructor to ensure that the machine is supplied with and has taken in the necessary ethical principles to both learn and perform the learned task in an ethical manner.

References

Anders, G. 1956/1979. *The Obsolescence of Man* (Die Antiquiertheit Des Menschen), 5th Edition. Munich: C. H. Beck.

Arnold, T., and M. Scheutz. 2017. "Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI", *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 445-452.

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.

Orseau, L. and Armstrong, S. 2016. "Safely Interruptible Agents". *In Proceeding of the Thirty-Second Conference on Uncertainty in Artificial* Intelligence, 557-566 .

Scheutz, M. 2014. "Teach One, Teach All". The 4th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent. http://ieeexplore.ieee.org/document/6917433/. (accessed March 1, 2018).

Scheutz, M. 2012. "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots". In *Anthology on Robot Ethics*, Patrick Lin and George Bekey and Keith Abney (eds.), MIT Press.

Wang, Y., and M. Kosinski. 2017. Deep Neural Networks Can Detect Sexual Orientation from Faces. *J. Personality Soc. Psychol.* **114**:246–257.

Williams, T., Briggs, P., and Scheutz, M. 2015. "Covert Robot-Robot Communication: Human Perceptions and Implications for HRI". *Journal of Human-Robot Interaction*, 4, 2, 23-49.