18

Ethical Aspects and Challenges for Interactive Task Learning

Matthias Scheutz

Abstract

As with all transformative technologies, humanity needs to analyze the ethical challenges and potential impacts associated with implementation. This chapter explores fundamental questions that pertain to interactive task learning (ITL): What is being taught and what are the associated risks? What are the dynamics of human–machine instruction? What effects will ITL have on human instructors and society? It explores the long-term impact that ITL could have on humans and human society, and discusses concerns valid to both machine learning and ITL (e.g., 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). Importantly, it stresses the unique aspects of ITL and proposes that the time to think about and take action on these concerns is now.

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, not the least due to their potential to bring significant benefits to 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 provides an example of a recent, potentially transformative technology—one 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?
- How can we guarantee that the machine will not behave in unethical ways after learning?

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

These challenges arise with all forms of machine learning, and thus with interactive task learning (ITL) as well. The difference with ITL, however, is that in addition to the ethical issues that pertain to the machine's newly acquired knowledge and how it will be used by the machine, questions arise regarding the learning interaction and how this interaction might subsequently affect humans. Moreover, there are questions about acquisition and use of the "normative surroundings" inherent to the task; that is, 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.

ITL offers many important advantages for humans and machines alike, especially when task-based natural language interactions are involved. For one, natural language instruction is an intuitive modality for humans, because humans are used to teaching each other through natural language dialogues. This means that less training and preparation are required on the human side (compared to implementing 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 agents 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, particularly for humans. We need to analyze, over time, what long-term impacts might result as humans engage with autonomous machines in ways that lead to successful knowledge acquisition for the machine. Some effects might be limited to the teaching interaction itself, whereas others could extend far beyond this context. Here, we examine three classes of ethical aspects that 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?

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 a risk that its learning process will not work as intended by the designers of the learning algorithm. The learning algorithm might:

- not acquire the right kind of information,
- only acquire incomplete knowledge, or
- acquire more knowledge than intended.

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

Consider, for example, the situation where a machine learning algorithm 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 (as in the first case); it may learn to detect faces but miss some faces under particular lighting conditions (the second case); or it could learn to detect faces as well as additional factors about humans such as their sexual orientation (the third case) (Wang and Kosinski 2017). Clearly, the result of unintended effects of machine learning could have ethical implications; for example, when unintended and unexpected information is exposed, misuse is possible, such as likely medical conditions or when wrongly classified information is used to make decisions that impact human lives (e.g., denying credit loan applications).

Learning algorithms in ITL are not exempt from these concerns regarding unintended consequences. Indeed, with ITL, additional aspects come into play that are typically not an issue with statistical machine learning from data. This is because ITL involves direct personal interactions between human instructors and machines, which is quite different from impersonal machine learning from data sets. For instance, if a data set does not contain information that the machine is supposed to learn, the data-driven machine learning algorithm is not at fault if the machine fails to extract it. In ITL, however, it is not clear who would be to blame if critical aspects of a task were 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 tutor 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 the blame would be put onto the machine (at least initially while ITL technology is still developing), the very possibility that a human instructor could get blamed (or would personally take on the blame) 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 humans involved.

Based on empirical work in human–robot interaction (Fan et al. 2017; Phillips et al. 2017), it seems reasonable to assume that cues of human likeness (e.g., natural language understanding, humanoid physical robots, virtual avatars) will cause people to import a vast array of assumptions about the

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

machine's capabilities that may not be warranted, such as background knowledge and level of understanding. 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 humanlike 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) which in turn could lead to unintended consequences during task performance: the machines, for instance, 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 a problem for a human learner who can "think along" and automatically correct such errors (Thomaz et al., this volume). 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 a human 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. Take, for example, 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 noninjured 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. Yet 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) 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, the machine must obtain the ingredients legally and must not simply take them from the surrounding areas (e.g., dismantling a neighbor's fence). Another example would be that when a Thanksgiving

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

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 already know them, and we rely on their ability to apply what a "reasonable person" would do. This legal term contains a lot of commonsense knowledge and reasoning that we, however, cannot assume machines will automatically possess. Thus, in addition to determining what legal and moral aspects of a task the instructor needs explicitly to teach, there might also be social normative aspects that the system would have to know to operate safely in human environments (e.g., no quick movement in crowded human environments, not moving toward people while holding knives that are pointed at them). Of course, cultural differences in norms pose further challenges as robots will have to be aware of the cultural background of their interactants (e.g., directness is considered polite by Russians, whereas English and German natives prefer indirectness; see Wierzbicka 1985).

The Dynamics of Human–Machine Instruction

The fundamental difference between ITL and other task-learning algorithms (e.g., learning tasks from instructional videos) involves real-time personal interaction with a human instructor. The human element imposes constraints and requirements on the types of interactions that machines may conduct with humans: interactions need to be respectful of human normative expectations (e.g., politeness) as well as human cognitive abilities (e.g., creativity, anticipation) and limitations (e.g., memory decay, limited focus of attention). The opportunities and challenges presented by these human characteristics do not have to be addressed by traditional data-driven learning algorithms. For instance, it is critical for machines to exhibit appropriate demeanor to ensure that humans will not be offended, and thus unwilling to continue an interaction. Examples of insensitivity could include a machine repeatedly dropping comments like "easy enough" and "no problem," which might be construed by the human as downplaying 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. For example, if a robot needs a screwdriver at a certain point in the task and sees the human instructor 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 be unable to meet human expectations in terms of natural language understanding, speed of actions, timing of interactions, etc. They will simply not be advanced enough in their interaction capabilities, background and commonsense knowledge, and natural language understanding to learn in ways that humans would assume when instructing other humans (as discussed above, such expectations

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

come naturally to humans when machines are appear to be humanlike). As a result, teaching interactions will quickly devolve into unnatural and tedious exchanges for humans. It will thus 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 that different interaction patterns might have on human emotions: 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.

After ITL and machine capabilities have sufficiently advanced, being aware of human limitations will also become a critical component to preserve human dignity. Ignoring human cognitive limitations (e.g., attention span, ability to remain focused and concentrate, speed of natural language processing and information intake) can lead to dysfunctional interactions that would not serve either interactant well. For instance, a robot that anticipates human instruction after only a few words and starts to execute it proactively might confuse people, in the simplest case, or even cause anxiety 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. A robot's physical constraints or agent's capabilities may 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? Imagine the following situation: After receiving instructions to follow the steps in a human instruction manual to assemble a chest of drawers, the robot detects various shortcuts and alternative ways of grasping and assembling parts. Based on this new knowledge, the robot proceeds to complete the task at a much faster speed than humanly possible, yet before the job is done, the robot's steps do not seem to make sense to the human, who recognizes that the robot has deviated from the manual. By way of its physical capabilities, the robot could perform multiple steps in parallel (e.g., if it has several grippers it can use independently) or determine alternative ways of connecting parts to improve the stability of the chest (e.g., based on its own physics models or mental simulations). Such unexpected, superhuman performance by the robot might not only make the social interaction element with people uncomfortable and the human teacher's job of ensuring proper learning and performance harder, it could also leave psychological marks on the human that persist beyond the interaction, as we will discuss next.

Extended Effects of ITL on Human Instructors and Society

An important ethical aspect of new technologies is their long-term impact on humans and human society. While ITL shares some of the same long-term

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

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—there are also unique, long-term ethical aspects that directly relate to human nature that need to be addressed. 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. Will humans feel (possibly unnecessarily) responsible when machines do not manage to acquire a task properly? Will humans blame themselves, instead of the machine, when a machine fails at a task because they really cared about the machine's success? Alternatively, when the machine succeeds after repeated learning interactions, will its success prompt feelings of pride in the human and lead potentially to the establishment of unidirectional emotional bonds (e.g., Scheutz 2014), because the human feels a personal connection with the machine?

Conversely, will the human be shocked, put off, or worried when observing machines with "superhuman" 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 to determine the length of a piece of wood that needs to be cut whereas a robot could immediately cut it using its visual system to measure the correct length). Consider machines that can consult cloud-based databases while interacting with human teachers, acquiring all necessary background knowledge quickly, on the fly, before human instruction has even finished; or machines that may covertly exchange messages with other learning machines while being instructed and quickly learn new skills from those machines. 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 learned everything it needs to know by 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 would affect labor markets or the economy is anybody's guess. It seems reasonable to assume that first-hand experience of such superhuman performance by machines—the awe as well as the jealousy and inferiority we may feel when the machine rapidly perfects a skill—could have profound ramifications for how we, as humans, view ourselves. In fact, it could lead to what the philosopher Günther Anders (1956/1979) called the "Promethean Shame": the feeling of inadequacy that results from watching

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

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.

Discussion

As with all new technologies, it is important to weigh the advantages and disadvantages of ITL and to consider carefully the trade-offs and risks involved in allowing, or even requiring, humans to teach machines. In the context of the larger discussion about the utility and dangers of artificial intelligence (AI), ITL certainly shares the worries that have been expressed in regard to machine learning:

- How can we guarantee that learning machines will be safe for humanity?
- How can we ensure that ITL agents can be turned off if they evolve in a dangerous direction and everything else fails?

These topics, currently discussed under the moniker "Big Red Button" (i.e., the means to shut off deviant AI) apply to ITL in the same way as they apply to other learning methods. Different from variants of reinforcement learning, where machines have to be incentivized to let them be shut off (Orseau and Armstrong 2016), ITL allows, however, for the explicit instruction of ethical principles in conjunction with tasks—an opportunity, if paired with the right computational architecture, that 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 2017).

Of course, instructing ethical principles along with tasks puts the burden of 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?
- Even worse, what if the instructor has a malicious agenda and aims to instruct the machine how to engage in terrorism?
- How would the machine know that such a task is off limits?

Underpinning much of the discussion about ITL is a tacit assumption that both teacher and learner will be benevolent: a teacher will not instruct inappropriate tasks and a learner only has a human's best interests in mind or, at the very least, avoids malicious intent. Such assumptions, however, may not always be warranted despite best intentions: teachers unaware of task and environmental conditions could make instructions ethically problematic; instructions might be contradictory and conflict resolution unclear; teachers could have ulterior

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

motives to teach tasks incorrectly; systems could be 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 follow human instructions blindly is a recipe for disaster.

In addition to the challenges associated with ensuring the ethical behavior of instructible machines, ITL poses additional challenges due to the intrinsic involvement of human instructors. These challenges intersect closely with related discussions on the ethics of human–robot interaction. Aside from the potential detrimental effects of ITL on the human psyche that have been anticipated by philosophers of technology for decades, questions of ownership, responsibility, and allegiance posed by ITL must 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?
- Who will assume responsibility for the robot's actions?

There are currently no good answers to any of these questions, partly because the research communities in AI, robotics, and human–robot interaction are still very much focused on understanding and addressing the fundamental technical challenges raised by ITL. What the above discussion has 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.

Conclusion

As new research emerges to advance the ability of machines to learn interactively from human instructors, it is imperative to keep in mind the overarching ethical aspects that pertain to the ITL learning algorithms, the learning interaction between human teacher and machine learner, and the long-term effects of the interaction on the human, so that the result will be ITL machines that benefit 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, and other ethical principles involved in learning and performing the task. Thus, machines

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.

instructed by ITL have the advantage of being able to learn their task as well as when, where, and how the task is 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.

From "Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions," edited by K. A. Gluck and J. E. Laird. Strüngmann Forum Reports, vol. 26, J. R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-03882-9.