2 Embedding Values Into Autonomous Intelligent Systems

Society does not have universal standards or guidelines to help embed human norms or moral values into autonomous intelligent systems (AIS) today. But as these systems grow to have increasing autonomy to make decisions and manipulate their environment, it is essential they be designed to adopt, learn, and follow the norms and values of the community they serve, and to communicate and explain their actions in as transparent and trustworthy manner possible, given the scenarios in which they function and the humans who use them.

The conceptual complexities surrounding what "values" are make it currently difficult to envision AIS that have computational structures directly corresponding to values. However, it is a realistic goal to embed explicit norms into such systems, because norms can be considered instructions to act in defined ways in defined contexts. A community's network of norms as a whole is likely to reflect the community's values, and AIS equipped with such a network would therefore also reflect the community's values, even if there are no directly identifiable computational structures that correspond to values.

To address this need, our Committee has broken the broader objective of embedding values into these systems into three major goals:

1. Identifying the norms and eliciting the values of a specific community affected by AIS.
2. Implementing the norms and values of that community within AIS.
3. Evaluating the alignment and compatibility of those norms and values between the humans and AIS within that community.

2 Embedding Values Into Autonomous Intelligent Systems

Pursuing these three goals represents an iterative process that is contextually sensitive to the requirements of AIS, their purpose, and their users within a specific community. It is understood that there will be clashes of values and norms when identifying, implementing, and evaluating these systems (a state often referred to as "moral overload"). This is why we advocate for a stakeholder-inclusive approach where systems are designed to provide transparent signals (such as explanations or inspection capabilities) about the specific nature of their behavior to the various actors within the community they serve. While this practice cannot always eliminate the possible data bias present in many machine-learning algorithms, it is our hope that the proactive inclusion of users and their interaction with AIS will increase trust in and overall reliability of these systems.

Identifying Norms and Values for Autonomous Intelligent Systems

Issue:

Values to be embedded in AIS are not universal, but rather largely specific to user communities and tasks.

Background

If machines enter human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms. A necessary step in enabling machines to do so is to identify these norms. Whereas laws are formalized and therefore relatively easy to identify, social and moral norms are more difficult to ascertain, as they are expressed through behavior, language, customs, cultural symbols, and artifacts. Moreover, communities (from families to whole nations) differ to various degrees in the norms they follow. So embedding norms in AIS requires a clear delineation of the community in which AIS are to be deployed. Further, even within the same community, different types of AIS will demand different sets of norms. The relevant norms for self-driving vehicles, for example, will differ greatly from those for robots used in healthcare.

Candidate Recommendation

We acknowledge that generating a universal set of norms/values that is applicable for all autonomous systems is not realistic. Instead, we recommend to first identify the sets of norms that AIS need to follow in specific communities and for specific tasks. Empirical research involving multiple disciplines and multiple methods should investigate and document these numerous sets of norms and make them available for designers to implement in AIS.

Further Resources

This book describes some of the challenges of having a one-size-fits-all approach to embedding human values in autonomous systems: Wallach, Wendell and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.

2 Embedding Values Into Autonomous Intelligent Systems

Issue:

Moral overload – AIS are usually subject to a multiplicity of norms and values that may conflict with each other.

Background

An autonomous system is often built with many constraints and goals in mind. These include legal requirements, monetary interests, and also social and moral values. Which constraints should designers prioritize? If they decide to prioritize social and moral norms of end users (and other stakeholders), how would they do that?

Candidate Recommendation

Our recommended best practice is to prioritize the values that reflect the shared set of values of the larger stakeholder groups. For example, a self-driving vehicle's prioritization of one factor over another in its decision making will need to reflect the priority order of values of its target user population, even if this order is in conflict with that of an individual designer, manufacturer, or client. For example, the Common Good Principle[vii] could be used as a guideline to resolve differences in the priority order of different stakeholder groups.

We also recommend that the priority order of values considered at the design stage of autonomous systems have a clear and explicit rationale. Having an explicitly stated rationale for value decisions, especially when these values are in conflict with one another, not only encourages the designers to reflect on the values being implemented in the system, but also provides a grounding and a point of reference for a third party to understand the thought process of the designer(s). The Common Good Principle mentioned above can help formulate such rationale.

We also acknowledge that, depending on the autonomous system in question, the priority order of values can dynamically change from one context of use to the next, or even within the same system over time. Approaches such as interactive machine learning (IML), or direct questioning and modeling of user responses can be employed to incorporate user input into the system. These techniques could be used to capture changing user values.

Further Resources

- Markkula Center for Applied Ethics, The Common Good. Idea of the common good decision-making was introduced here.
- Van den Hoven, Jeroen, Engineering and the Problem of Moral Overload. *Science and Engineering Ethics* 18, no. 1 (March 2012): 143-155.
- One of the places where differences in human moral decision-making and changes in priority order of values for autonomous systems are documented is a series of poll results published by the Open Roboethics initiative. In particular, see these poll results on care robots.

2 Embedding Values Into Autonomous Intelligent Systems

Issue:

AIS can have built-in data or algorithmic biases that disadvantage members of certain groups.

Background

Autonomous intelligent systems, compared to traditional systems, are sometimes discussed as a new type of species—called the new ontological category,[vii] according to literature in human-robot interaction—because of the manner in which humans perceive, interact with, and psychologically respond to them. For example, numerous studies have documented the way in which humans willingly follow even the strangest of requests from a robot, demonstrating the impact these systems can have on our decision-making and behavior (see for example, Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner, "Overtrust of Robots in Emergency Evacuation Scenarios,"[viii] 2016 ACM/IEEE International Conference on Human-Robot Interaction). Hence, it is important to be aware of possible use of the systems for the purposes of manipulation.

In addition, various aspects of these systems can be designed to instill bias into other users, whether intended or not. The sources of bias can span from the way a system senses the world (e.g., can the system detect a person missing an arm or does it assume all humans have two arms?), to how it processes and responds to the sensed information (e.g., does the system respond to people of different ethnicity, gender, race, differently?), as well as what it looks like. Details of an interactive autonomous system's behavior can have far-reaching consequences, such as reinforcement of gender, ethnic, and other biases (see for example, Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,"[ix] Cornell University Library, arXiv:1607.06520, July 21, 2016.)

Moreover, while deciding which values and norms to prioritize, we call for special attention to the interests of vulnerable and under-represented populations, such that these user groups are not exploited or disadvantaged by (possibly unintended) unethical design. While traditionally the term *vulnerable populations* refers to disadvantaged sub-groups within human communities—including but not limited to children, older adults, prisoners, ethnic minorities, economically disadvantaged, and people with physical or intellectual disabilities—here we also include populations who may not be traditionally considered a member of vulnerable populations, but may be so in the context of autonomous intelligent systems. For example, riders in autonomous vehicles, or factory workers using a 400-pound high-torque robot, who would not otherwise be vulnerable under the traditional definition, become vulnerable in the use contexts due to the user's reliance on the system or physical disadvantage compared to the high-powered machinery.

2 Embedding Values Into Autonomous Intelligent Systems

Candidate Recommendation

It is important to acknowledge that it is easy to have built-in biases in autonomous systems. For example, a system that depends on face recognition trained entirely on Caucasian faces may work incorrectly or not at all on people with non-Caucasian skin tones or facial structures. This renders the system to be perceived as discriminatory, whether it was designed with such intent or not. These biases can also stem from the values held by the designer. We can reduce the incidence of such unintended biases by being more aware of the potential sources of these biases. We posit that being aware of this particular issue and adopting more inclusive design principles can help with this process. For example, systems that can sense persons of different races, ethnicities, genders, ages, body shapes, or people who use wheelchairs or prosthetics, etc.

We also highlight that this concern delves into the domain of ongoing research in human-robot interaction and human-machine interaction. To what extent and how do built-in biases change the course of robot interaction with human users? What dynamic and longitudinal effect do they have on the users and the society? How does a robot's morphology in different use cases affect target user groups? These are all open research questions for which we do not yet have clear answers. Since there is no clear understanding of the nature of these biases and their alignment with human values, we recommend conducting research and educational efforts to resolve these open questions and to address these issues in a participatory way by introducing into the design process members of the groups who may be disadvantaged by the system.

In particular, vulnerable populations are often one of the first users of autonomous systems. In designing for these populations, we recommend designers familiarize themselves with relevant resources specific to the target population. We also note that a system can have multiple end users, each of which may demand a conflicting set of values. We recommend designers be aware of such conflicts and be transparent in addressing these conflicting value priorities as suggested in the above-mentioned issue. AIS are usually subject to a multiplicity of norms and values that may conflict with each other.

Therefore, we strongly encourage the inclusion of intended stakeholders in the entire engineering process, from design and implementation to testing and marketing, as advocated for example in disability studies literature (see "Nothing About Us Without Us" in the Further Resources below).

A number of institutions have established connections with communities of a particular vulnerable population (e.g., University of Washington's DO-IT program). However, there is no one voice that represents all vulnerable populations. Hence, we recommend designers and practitioners reach out to communities of interest and relevant advocacy groups.

We also recommend, especially when designing for dynamically vulnerable populations, that designers take on an interdisciplinary approach and involve relevant experts or advisory group(s) into the design process. Thus, designers of AIS should work together with behavioral scientists

2 Embedding Values Into Autonomous Intelligent Systems

and members of the target populations to systematically study population norms, expectations, concerns, and vulnerabilities. We also encourage designers to include regulators and policymakers in this process as well, noting that shaping regulation and policy is an integral part of guiding the development and deployment of autonomous systems in a desirable direction.

Further Resources

- Asaro, P. "Will BlackLivesMatter to RoboCop?" We Robot, 2016.
- Riek, L. D. and D. Howard. A Code of Ethics for the Human-Robot Interaction Profession. We Robot, 2014.
- Winfield, A. Robots Should Not Be Gendered (blog), 2016.
- Whitby, Blay. "Sometimes It's Hard to Be a Robot: A Call for Action on the Ethics of Abusing Artificial Agents." *Interacting with Computers* 20, no. 3 (2008): 326-333.
- Federal Trade Commission. Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress. 2000.
- Riek, Laurel D. "Robotics Technology in Mental Health Care." *Artificial Intelligence in Behavioral Health and Mental Health Care*, (2015): 185-203.
- Charlton, James I. *Nothing About Us Without Us: Disability Oppression and Empowerment*, University of California Press, 2000.
- Shivayogi, P. "Vulnerable Population and Methods for Their Safeguard." *Perspectives in Clinical Research*, January-March (2013): 53-57.

Embedding Norms and Values in Autonomous Intelligent Systems

Issue:

Once the relevant sets of norms (of AIS's specific role in a specific community) have been identified, it is not clear how such norms should be built into a computational architecture.

Background

The prospect of developing computer systems that are sensitive to human norms and values and factoring these issues into making decisions in morally or legally significant situations has intrigued science fiction writers, philosophers, and computer scientists alike. Modest efforts to realize this worthy goal in limited or bounded contexts are already underway. This emerging field of research goes under many names including: machine morality, machine ethics, moral machines, value alignment, computational ethics, artificial morality, safe AI, and friendly AI. Basic notions can be found in books such as Allen, C., and W. Wallach. *Moral Machines.*[x] *Teaching Robots Right from Wrong*. Oxford University Press, 2010.

Computers and robots already instantiate values in their choices and actions, but these values are programmed or designed by the engineers that build the systems. Increasingly, autonomous systems will encounter situations that their designers cannot anticipate, and will require algorithmic procedures to select the better of two or more possible courses of action. Some of the existing experimental approaches to building moral machines are top-down. In this sense the norms, rules, principles, or procedures are used by the system to evaluate the acceptability of differing courses of action or as moral standards or goals to be realized.

Recent breakthroughs in machine learning and perception will enable researchers to explore bottom-up approaches—in which the AI system learns about its context and about human values—similar to the manner in which a child slowly learns which forms of behavior are safe and acceptable. Of course a child can feel pain and pleasure, empathize with others, and has other capabilities that AI system cannot presently imitate. Nevertheless, as research on autonomous systems progresses, engineers will explore new ways to either simulate these capabilities, or build alternative mechanisms that fulfill similar functions.

2 Embedding Values Into Autonomous Intelligent Systems

Candidate Recommendation

Research on this front should be encouraged. Advances in data collection, sensor technology, pattern recognition, machine learning, and integrating different kinds of data sets will enable creative, new approaches for ensuring that the actions of AI systems are aligned with the values of the community in which they operate. Progress toward building moral machines may well determine the safety and trustworthiness of increasingly autonomous AI systems.

Further Resources

- Allen, C., and W. Wallach. *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, 2010.
- Anderson, M., and S. Anderson (eds.). *Machine Ethics.* Cambridge University Press, 2011.
- Abney, K., G. Bekey, and P. Patrick. *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT Press, 2011.
- RC Arkin, P Ulam, AR Wagner, Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception, Proceedings of the IEEE 100 (3), 571-589

Evaluating the Alignment of Norms and Values between Humans and AIS

Issue:

Norms implemented in AIS must be compatible with the norms in the relevant community.

Background

If a community's systems of norms (and their underlying values) has been identified, and if this process has successfully guided the implementation of norms in AIS, then the third step in value embedding must take place: rigorous testing and evaluation of the resulting human-machine interactions regarding these norms.

An intuitive criterion in these evaluations might be that the norms embedded in AIS should correspond closely to the human norms identified in the community—that is, AIS should be disposed to behave the same way that people expect each other to behave. However, for a given community and a given AIS task and use context, AIS and humans may not have identical, but rather compatible, sets of norms. People will have some unique expectations for humans that they don't have for machines (e.g., norms governing the expression of emotions, as long as machines don't have, or clearly express, emotions), and people will have some unique expectations of AIS that they don't have for humans (e.g., that the machine will destroy itself if it can thereby prevent harm to a human). The norm identification process must document these structural relations (similarities as well as differences) between human and AIS norms, and in evaluating these relations, the goal of *compatibility* may be preferred over that of *alignment*, which suggests primarily a similarity structure.

In addition, more concrete criteria must be developed that indicate the quality of human-machine interactions, such as human approval and appreciation of AIS, trust in AIS, adaptability of AIS to humans users, and human benefits in the presence or influence of AIS. Evaluation of these and other criteria must occur both before broad deployment and throughout the life cycle of the system. Assessment before deployment would best take place in systematic test beds that allow human users (from the defined community) to engage safely with AIS (in the defined tasks) and enable assessment of approval, trust, and related variables. Examples include the Tokku testing zones in Japan.[xi]

2 Embedding Values Into Autonomous Intelligent Systems

Candidate Recommendation

The success of implementing norms in AIS must be rigorously evaluated by empirical means, both before and throughout deployment. Criteria of such evaluation will include compatibility of machine norms and human norms (so-called *value alignment* or *compliance*, depending on the nature of the norms), human approval of AIS, and trust in AIS, among others. Multiple disciplines and methods should contribute to developing and conducting such evaluation, such as extensive tests (including adversarial ones), explanation capabilities to reconstruct AIS inner functioning, natural language dialog between AIS and humans (including deep question answering), and context awareness and memory (to handle repeated evaluations).

Issue:

Achieving a correct level of trust between humans and AIS.

Background

Development of autonomous systems that are worthy of our trust is challenged due to the current lack of transparency and verifiability regarding these systems for users. For this issue, we explore two levels at which transparency and verifiability are useful and often necessary. A first level of transparency relates to the information conveyed to the user while an autonomous system interacts with the user. A second level has to do with the possibility to evaluate the system as a whole by a third party (e.g., regulators, society at large, and post-accident investigators).

In the first level, consider for example the case of robots built to interact with people. The robots should be designed to be able to communicate what they are about to perform and why as the actions unfold. This is important in establishing an appropriate level of trust with the user. While a system that a user does not trust may never be used, a system that is overly trusted can negatively affect the user as well based on the perception of the particular system or similar types of systems by the society. Unlike humans who naturally use verbal and nonverbal behaviors to convey trust-based information to those around them, the mode and the content of communicative behaviors toward or from an autonomous system are features that would be absent if not for the explicit implementation by the designers. Designing systems that are worthy of our trust necessarily includes making these explicit design decisions. As with people, trust is built over time, through repeated interactions, so AIS must be equipped with context awareness and memory capabilities.

Candidate Recommendation

Transparency and verifiability are necessary for building trust in AIS. We recommend that AIS come equipped with a module assuring some level of transparency and verifiability. Technological solutions to address the issue of transparency and instilling the right level of trust in the users is an open area of research. Trust

2 Embedding Values Into Autonomous Intelligent Systems

is also a dynamic variable in human-machine interaction; the level of trust a user may have with a system tends to change over time. Coupled with the dynamic nature of trust in autonomous systems is our known tendency to overly trust technology beyond its capabilities. With systems that have been commercialized, for example, users often assume a minimum level of reliability and trustworthiness of the system from the onset.

Hence, even when a system is delivered with a written disclaimer outlining its conditions of use, it is often naïve to assume that the disclaimer alone can protect the interests of both the manufacturer/developer and users. In addition to communicating the limitations and capabilities of the system to the users, we recommend autonomous systems to be designed with features that prevent users from operating the system outside a known, safe, and appropriate range of conditions of use, including conditions that depend on user behavior. We also recommend evaluation of the system's design with the user's perception of their role in mind (e.g., operator versus user of the system), such that the system's interaction with the user is in alignment with the role that is expected of the user.

In addition, one can design communicative and behavioral features of a system to serve as interactive real-time disclaimers, such that the user is informed of significant changes to the system's level of confidence on a proposed solution for the task to be performed, which can change from one moment or situation to the next. Systems that lack such features can result in not only ineffective interaction with the user—introducing a point of miscommunication, for example—but also risk the safety and wellbeing of the user and others. This also makes it more challenging for a user to diagnose the reasons why a system may be behaving in a certain way, and to detect when malfunctions occur.

Issue:

Third-party evaluation of AIS's value alignment.

Background

The second level of transparency, as stated above, is needed to evaluate a system as a whole by a third party (e.g., regulators, society at large, and post-accident investigators).

In this second category, there are concerns regarding the increasing number of autonomous systems that rely on, or include, AI/machine-learning techniques inherently lacking transparency and verifiability. Discussions on this topic include: the nature and possible bias of the data sets used to train a machine-learning system that is often not accessible by the public, details of the algorithm used to create the final product, the specifications on the final product's efficacy and performance, and the need to consider the scenario where AIS will be used when evaluating their adherence to relevant human values. While acknowledging the usefulness and potential for

2 Embedding Values Into Autonomous Intelligent Systems

these systems, it is a serious concern that even the designers and programmers involved cannot verify or guarantee reliability, efficacy, and value alignment of the final system. A further problem is that there is no agreed-upon method, process, or standards for validating and certifying the adherence of AIS to desired human norms and values.

Candidate Recommendation

With regards to our concern on the transparency between a system as a whole and its evaluator (e.g., regulator), we recommend that designers and developers alike document changes to the systems in their daily practice. A system with the highest level of traceability would contain a black-box-like module such as those used in the airline industry, that logs and helps diagnose all changes and behaviors of the system. Such practice, while it does not fully address the need for transparency of a number of popular machine-learning approaches, allows one to trace back to the sources of problems that may occur and provide a mechanism with which a faulty behavior of a system can be diagnosed.

As more human decision-making is delegated to autonomous systems, we expect there to be an increasing need for rationale and explanation as to how the decision was reached by the algorithm. In this respect, a relevant regulation is the European Union's new General Data Protection Regulation (GDPR)[xiv], adopted on April 2016 and scheduled to take effect in 2018. The GDPR states that, in regards to automated decisions based on personal data, individuals have a right to "an explanation of the [algorithmic] decision reached after such assessment and to challenge the decision." While the development of an algorithm that is able to explain its behavior is an open research topic, there are algorithms that are more transparent than others, such as logic-based AI that provide more transparency than machine-learning AI, and more coherence between the output behavior of a system and its inner functioning. Winfield, Blum, and Liu's work on consequence engine[xv], for example, utilizes a simulator to predict and evaluate the consequences of an artificial agent's possible next actions in order to decide the right course of action, making the agent's decision-making process easy to examine and validate. In the absence of an adequate alternative, it is imperative that designers be aware of the need for transparency and strive to increase it in the algorithms they design and implement into autonomous systems.

We also recommend that regulators define, together with users, developers, and designers, a minimum level of value alignment and compliance, and suitable capabilities for this to be checked by a third party, in order for AIS to be deployed.

Finally, we recommend to define criteria to define AIS as trustworthy. These criteria will depend on a machine's expected tasks and context of use, as well as the users' vulnerabilities (we expect that more-vulnerable-user categories will require more stringent criteria).

2 Embedding Values Into Autonomous Intelligent Systems

Further Resources

- Goodman, B., and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'," Cornell University Library, arXiv: 1606.08813, August 31, 2016.
- Winfield, A. F. T., C. Blum, and W. Liu, "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection," *Advances in Autonomous Robotics Systems*, Lecture Notes in Computer Science Volume 8717, (2014): 85-96. Eds. Mistry M, Leonardis A, Witkowski M and Melhuish C, Springer, 2014.