Embedding Values into Autonomous and Intelligent Systems

Society has not established universal standards or guiding principles for embedding human values and norms into autonomous and intelligent systems (A/IS) today. But as these systems are instilled with increasing autonomy in making decisions and manipulating their environment, it is essential that they are designed to adopt, learn, and follow the norms and values of the community they serve. Moreover, their actions should be transparent in signaling their norm compliance and, if needed, they must be able to explain their actions. This is essential if humans are to develop appropriate levels of trust in A/IS in the specific contexts and roles in which A/IS function.

At the present time, the conceptual complexities surrounding what "values" are (Hitlin and Piliavin 2004[1]; Malle and Dickert 2007[2]; Rohan 2000[3]; Sommer 2016[4]) make it difficult to envision A/IS that have computational structures directly corresponding to social or cultural values such as "security," "autonomy," or "fairness". It may be a more realistic goal to embed explicit norms into such systems. Since norms are observable in human behavior, they can therefore be represented as instructions to act in defined ways in defined contexts, for a specific community—from family to town to country and beyond. A community's network of social and moral norms is likely to reflect the community's values, and A/IS equipped with such a network would, therefore, also reflect the community's values. For discussion of specific values that are critical for ethical considerations of A/IS, see the chapters of *Ethically Aligned Design,* "Personal Data and Individual Agency" and "Well-being".

Norms are typically expressed in terms of obligations and prohibitions, and these can be expressed computationally (Malle, Scheutz, and Austerweil 2017[5]; Vázquez-Salceda, Aldewereld and Dignum 2004[6]). They are typically qualitative in nature, e.g., do not stand too close to people. However, the implementation of norms also has a quantitative component—the measurement of the physical distance we mean by "too close", and the possible instantiations of the quantitative component technically enable the qualitative norm.

To address the broad objective of embedding norms and, by implication, values into A/IS, this chapter addresses three more concrete goals:

1. Identifying the norms of the specific community in which the A/IS operate,
2. Computationally implementing the norms of that community within the A/IS, and
3. Evaluating whether the implementation of the identified norms in the A/IS are indeed conforming to the norms reflective of that community.

Pursuing these three goals represents an iterative process that is sensitive to the purpose of the A/IS and to its users within a specific community. It is understood that there may be conflicts of values and norms when identifying, implementing, and evaluating these systems. Such conflicts are a natural part of the dynamically changing and renegotiated norm systems of any community. As a result, we advocate for an approach in which systems are designed to provide transparent signals describing the specific nature of their behavior to the individuals in the community they serve. Such signals may include explanations or offers for inspection and must be in a language or form that is meaningful to the community.

Further Resources

- S. Hitlin and J. A. Piliavin, "Values: Reviving a Dormant Concept." *Annual Review of Sociology* 30, pp.359–393, 2004.
- B. F. Malle, and S. Dickert. "Values," in *Encyclopedia of Social Psychology*, edited by R. F. Baumeister and K. D. Vohs. Thousand Oaks, CA: Sage, 2007.
- B. F. Malle, M. Scheutz, and J. L. Austerweil. "Networks of Social and Moral Norms in Human and Robot Agents," in A World with Robots: *International Conference on Robot Ethics*: ICRE 2015, edited by M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, and G. S. Virk, 3–17. Cham, Switzerland: Springer International Publishing, 2017.
- M. J. Rohan, "A Rose by Any Name? The Values Construct." *Personality and Social Psychology Review* 4, pp. 255–277, 2000.
- U. Sommer, Werte: *Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt.* [Values. Why We Need Them Even Though They Don't Exist.] Stuttgart, Germany: J. B. Metzler, 2016.
- J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. "Implementing Norms in Multiagent Systems," in Multiagent System Technologies. MATES 2004, edited by G. Lindemann, Denzinger, I. J. Timm, and R. Unland. (Lecture Notes in Computer Science, vol. 3187.) Berlin: Springer, 2004.

Section 1—Identifying Norms for Autonomous and Intelligent Systems

We identify three issues that must be addressed in the attempt to identify norms and corresponding values for A/IS. The first issue asks which norms should be identified and with which properties. Here we highlight context specificity as a fundamental property of norms. Second, we emphasize another important property of norms: their dynamically changing nature (Mack 2018[7]), which requires A/IS to have the capacity to update their norms and learn new ones. Third, we address the challenge of norm conflicts that naturally arise in a complex social world. Resolving such conflicts requires priority structures among norms, which help determine whether, in a given context, adhering to one norm is more important than adhering to another norm, often in light of overarching standards, e.g., laws and international humanitarian principles.

Issue 1: Which norms should be identified?

Background

If machines engage in human communities, 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. But which norms should be identified? Laws are publicly documented and therefore easy to identify, so they can be incorporated into A/IS as long as they do not violate humanitarian or community moral principles. Social and moral norms are more difficult to ascertain, as they are expressed through behavior, language, customs, cultural symbols, and artifacts. Most important, communities ranging from families to whole nations differ to various degrees in the norms they follow. Therefore, generating a universal set of norms that applies to all A/IS in all contexts is not realistic, but neither is it advisable to completely tailor the A/IS to individual preferences. We suggest that it is feasible to identify broadly observed norms of communities in which a technology is deployed.

Furthermore, the difficulty of generating a universal set of norms is not inconsistent with the goal of seeking agreement over Universal Human Rights (see the "General Principles" chapter of *Ethically Aligned Design*). However, these universal rights are not sufficient for devising A/IS that conform to the specific norms of its community. Universal Human Rights must, however, constrain the kinds of norms that are implemented in the A/IS (cf. van de Poel 2016[8]).

Embedding norms in A/IS requires a careful understanding of the communities in which the A/IS are to be deployed. Further, even within a particular community, different types of A/IS will demand different sets of norms. The relevant norms for self-driving vehicles, for example, may differ greatly from those for robots used in healthcare. Thus, we recommend that to develop A/IS capable of following legal, social, and moral norms, the first step is to identify the norms of the specific community in which the A/IS are to be deployed and, in particular, norms relevant to the kinds of tasks and roles for which the A/IS are designed. Even when designating a narrowly defined community, e.g., a nursing home, an apartment complex, or a company, there will be variations in the norms that apply, or in their relative weighting. The norm identification process must heed such variation and ensure that the identified norms are representative, not only of the dominant subgroup in the community but also of vulnerable and underrepresented groups.

The most narrowly defined "community" is a single person, and A/IS may well have to adapt to the unique expectations and needs of a given individual, such as the arrangement of a disabled person's living accommodations. However, unique individual expectations must not violate norms in the larger community. Whereas the arrangement of someone's kitchen or the frequency with which a care robot checks in with a patient can be personalized without violating any community norms, encouraging the robot to use derogatory language to talk about certain social groups does violate such norms. In the next section, we discuss how A/IS might handle such norm conflicts.

Innovation projects and development efforts for A/IS should always rely on empirical research, involving multiple disciplines and multiple methods; to investigate and document both context- and task-specific norms, spoken and unspoken, that typically apply in a particular community. Such a set of empirically identified norms should then guide system design. This process of norm identification and implementation must be iterative and revisable. A/IS with an initial set of implemented norms may betray biases of original assessments (Misra, Zitnick, Mitchell, and Girshick 2016[9]) that can be revealed by interactions with, and feedback from, the relevant community. This leads to a process of norm updating, which is described next in Issue 2.

Recommendation

To develop A/IS capable of following social and moral norms, the first step is to identify the norms of the specific community in which the A/IS are to be deployed and, in particular, norms relevant to the kinds of tasks and roles that the A/IS are designed for. This norm identification process must use appropriate scientific methods and continue through the system's life cycle.

Further Resources

- Mack, Ed., "Changing social norms." *Social Research: An International Quarterly,* 85, no.1, 1–271, 2018.
- I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, (2016). Seeing through the human reporting bias: Visual Classifiers from Noisy Human-Centric Labels. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 2930–2939. doi:10.1109/CVPR.2016.320
- I. van de Poel, "An Ethical Framework for Evaluating Experimental Technology," *Science and Engineering Ethics*, 22, no. 3,pp. 667-686, 2016.

Issue 2: The need for norm updating

Background

Norms are not static. They change over time, in response to social progress, political change, new legal measures, or novel opportunities (Mack 2018[10]). Norms can fade away when, for whatever reasons, fewer and fewer people adhere to them. And new norms emerge when technological innovation invites novel behaviors and novel standards, e.g., cell phone use in public.

A/IS should be equipped with a starting set of social and legal norms before they are deployed in their intended community (see Issue 1), but this will not suffice for A/IS to behave appropriately over time. A/IS or the designers of A/IS, must be adept at identifying and adding new norms to its starting set, because the initial norm identification process in the community will undoubtedly have missed some norms and because the community's norms change.

Humans rely on numerous capacities to update their knowledge of norms and learn new ones. They observe other community members' behavior and are sensitive to collective norm change; they explicitly ask about new norms when joining new communities, e.g., entering college or a job in a new town; and they respond to feedback from others when they exhibit uncertainty about norms or have violated a norm.

Likewise, A/IS need multiple capacities to improve their own norm knowledge and to adapt to a community's dynamically changing norms. These capacities include:

- Processing behavioral trends by members of the target community and comparing them to trends predicted by the baseline norm system,
- Asking for guidance from the community when uncertainty about applicable norms exceeds a critical threshold,
- Responding to instruction from the community members who introduce a robot to a previously unknown context or who notice the A/IS' uncertainty in a familiar context, and
- Responding to formal or informal feedback from the community when the A/IS violate a norm.

The modification of a normative system can occur at any level of the system: it could involve altering the priority weightings between individual norms, changing the qualitative expression of a norm, or altering the quantitative parameters that enable the norm.

We recommend that the system's norm changes be transparent. That is, the system or its designer should consult with users, designers, and community representatives when adding new norms to its norm system or adjusting the priority or content of existing norms. Allowing a system to learn new norms without public or expert review has detrimental consequences (Green and Hu 2018[11]). The form of consultation and the specific review process will vary by machine sophistication—e.g., linguistic capacity and function/role, or a flexible social companion versus a task-defined medical robot—and best practices will have to be established. In some cases, the system may document its dynamic change, and the user can consult this documentation as desired. In other cases, explicit announcements and requests for discussion with the designer may be appropriate. In yet other cases, the A/IS may propose changes, and the relevant human community, e.g., drawn from a representative crowdsourced panel, will decide whether such changes should be implemented in the system.

Recommendation

To respond to the dynamic change of norms in society A/IS or their designers must be able to amend their norms or add new ones, while being transparent about these changes to users, designers, broader community representatives, and other stakeholders.

Further Resources

- B. Green and L. Hu. "The Myth in the Methodology: Towards a Recontextualization of Fairness in ML." Paper presented at the Debates workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden 2018.
- Mack, Ed., "Changing social norms," *Social Research: An International Quarterly*, 85 (1, Special Issue), 1-271, 2018.

Issue 3: A/IS will face norm conflicts and need methods to resolve them.

Background

Often, even within a well-specified context, no action is available that fulfills all obligations and prohibitions. Such situations—often described as moral dilemmas or moral overload (Van den Hoven 2012[12])—must be computationally tractable by A/IS; they cannot simply stop in their tracks and end on a logical contradiction. Humans resolve such situations by accepting trade-offs between conflicting norms, which constitute priorities of one norm or value over another in a given context. Such priorities may be represented in the norm system as hierarchical relations.

Along with identifying the norms within a specific community and task domain, empirical research must identify the ways in which people prioritize competing norms and resolve norm conflicts, and the ways in which people expect A/IS to resolve similar norm conflicts. These more local conflict resolutions will be further constrained by some general principles, such as the "Common Good Principle" (Andre and Velasquez 1992[13]) or local and national laws. For example, a self-driving vehicle's prioritization of one factor over another in its decision-making will need to reflect the laws and norms of the population in which the A/IS are deployed, e.g., the traffic laws of a U.S. state and the United States as a whole.

Some priority orders can be built into a given norm network as hierarchical relations, e.g., more general prohibitions against harm to humans typically override more specific norms against lying. Other priority orders can stem from the override that norms in the larger community exert on norms and preferences of an individual user. In the earlier example discussing personalization (see Issue 1), the A/IS of a racist user who demands the A/IS use derogatory language for certain social groups will have to resist such demands because community norms hierarchically override an individual user's preferences. In many cases, priority orders are not built in as fixed hierarchies because the priorities are themselves context-specific or may arise from net moral costs and benefits of the particular case at hand. A/IS must have learning capacities to track such variations and incorporate user and community input, e.g., about the subtle differences between contexts, so as to refine the system's norm network (see Issue 2).

Tension may sometimes arise between a community's social and legal norms and the normative considerations of designers or manufacturers. Democratic processes may need to be developed that resolve this tension—processes that cannot be presented in detail in this chapter. Often such resolution will favor the local laws and norms, but in some cases the community may have to be persuaded to accept A/IS favoring international law or broader humanitarian principles over, say, racist or sexist local practices.

In general, we recommend that the system's resolution of norm conflicts be transparent—that is, documented by the system and ready to be made available to users, the relevant community of deployment, and third-party evaluators. Just like people explain to each other why they made decisions, they will expect any A/IS to be able to explain their decisions and be sensitive to user feedback about the appropriateness of the decisions. To do so, design and development of A/IS should specifically identify the relevant groups of humans who may request explanations and evaluate the systems' behaviors. In the case of a system detecting a norm conflict, the system should consult and offer explanations to representatives from the community, e.g., randomly sampled crowdsourced members or elected officials, as well as to third-party evaluators, with the goal of discussing and resolving the norm conflict.

Recommendation

A/IS developers should identify the ways in which people resolve norm conflicts and the ways in which they expect A/IS to resolve similar norm conflicts. A system's resolution of norm conflicts must be transparent—that is, documented by the system and ready to be made available to users, the relevant community of deployment, and third-party evaluators.

Further Resources

- M. Velasquez, C. Andre, T. Shanks, S.J., and M. J. Meyer, "The Common Good." *Issues in Ethics*, vol. 5, no. 1, 1992.
- J. Van den Hoven, "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics, vol.* 18, no. 1, pp. 143–155, 2012.
- D. Abel, J. MacGlashan, and M. L. Littman. "Reinforcement Learning as a Framework for Ethical Decision Making." *AAAI Workshop AI, Ethics, and Society, Volume WS-16-02 of 13th AAAI Workshops*. Palo Alto, CA: AAAI Press, 2016.
- O. Bendel, Die Moral in der Maschine: Beiträge zu Roboter- und Maschinenethik. Hannover, Germany: Heise Medien, 2016.
  - Accessible popular-science contributions to philosophical issues and technical implementations of machine ethics
- S. V. Burks, and E. L. Krupka. "A Multimethod Approach to Identifying Norms and Normative Expectations within a Corporate Hierarchy: Evidence from the Financial Services Industry." *Management Science,* vol. 58, pp. 203–217, 2012.
  - Illustrates surveys and incentivized coordination games as methods to elicit norms in a large financial services firm
- F. Cushman, V. Kumar, and P. Railton, "Moral Learning," *Cognition*, vol. 167, pp. 1–282, 2017.
- M. Flanagan, D. C. Howe, and H. Nissenbaum, "Embodying Values in Technology: Theory and Practice." *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, Eds., Cambridge University Press, 2008, pp. 322–53. Cambridge Core, *Cambridge University Press.* Preprint available at http://www.nyu.edu/projects/nissenbaum/papers/Nissenbaum-VID.4-25.pdf
- B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren. "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory,* N. Doorn, Schuurbiers, I. van de Poel, and M. Gorman, Eds., vol. 16, pp. 55–95. Dordrecht: Springer, 2013.
  - A comprehensive introduction into Value Sensitive Design and three sample applications
- G. Mackie, F. Moneti, E. Denny, and H. Shakya. "What Are Social Norms? How Are They Measured?" UNICEF Working Paper. University of California at San Diego: UNICEF, Sept. 2014. https://dmeforpeace.org/sites/default/files/4%2009%2030%20Whole%20What%20are%20Social%20Norms.pdf
  - A broad survey of conceptual and measurement questions regarding social norms.
- J. A. Leydens and J. C. Lucena. Engineering Justice: Transforming Engineering Education and Practice. Hoboken, NJ: John Wiley & Sons, 2018.
  - Identifies principles of engineering for social justice.

- B. F. Malle, "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology,* vol. 18, no. 4, pp. 243–256, 2016.
    - Discusses how a robot's norm capacity fits in the larger vision of a robot with moral competence.
- K. W. Miller, M. J. Wolf, and F. Grodzinsky, "This 'Ethical Trap' Is for Roboticists, Not Robots: On the Issue of Artificial Agent Ethical Decision-Making." *Science and Engineering Ethics,* vol. 23, pp. 389–401, 2017.
    - This article raises doubts about the possibility of imbuing artificial agents with morality, or of claiming to have done so.
- Open Roboethics Initiative: www.openroboethics.org. A series of poll results on differences in human moral decision-making and changes in priority order of values for autonomous systems (e.g., on care robots), 2019.
- A. Rizzo and L. L. Swisher, "Comparing the Stewart–Sprinthall Management Survey and the Defining Issues Test-2 as Measures of Moral Reasoning in Public Administration." *Journal of Public Administration Research and Theory,* vol. 14, pp. 335–348, 2004.
    - Describes two assessment instruments of moral reasoning (including norm maintenance) based on Kohlberg's theory of moral development.
- S. H. Schwartz, "An Overview of the Schwartz Theory of Basic Values." *Online Readings in Psychology and Culture* 2, 2012.
    - Comprehensive overview of a specific theory of values, understood as motivational orientations toward abstract outcomes (e.g., self-direction, power, security).
- S. H. Schwartz and K. Boehnke. "Evaluating the Structure of Human Values with Confirmatory Factor Analysis." *Journal of Research in Personality,* vol. 38, pp. 230–255, 2004.
    - Describes an older method of subjective judgments of relations among valued outcomes and a newer, formal method of analyzing these relations.
- W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.
    - This book describes some of the challenges of having a one-size-fits-all approach to embedding human values in autonomous systems.

Section 2—Implementing Norms in Autonomous and Intelligent Systems

Once the norms relevant to A/IS' role in a specific community have been identified, including their properties and priority structure, we must link these norms to the functionalities of the underlying computational system. We discuss three issues that arise in this process of norm implementation. First, computational approaches to enable a system to represent, learn, and execute norms are only slowly emerging. However, the diversity of approaches may soon lead to substantial advances. Second, for A/IS that operate in human communities, there is a particular need for transparency—ranging from the technical process of implementation to the ethical decisions that A/IS will make in human-machine interactions, which will require a high level of explainability. Third, failures of normative reasoning can be considered inevitable and mitigation strategies should therefore be put in place to handle such failures when they occur.

As a general guideline, we recommend that, through the entire process of implementation of norms, designers should consider various forms and metrics of evaluation, and they should define and incorporate central criteria for assessing the A/IS' norm conformity, e.g., human-machine agreement on moral decisions, verifiability of A/IS decisions, or justified trust. In this way, implementation already prepares for the critical third phase of evaluation (discussed in Section 3).

Issue 1: Many approaches to norm implementation are currently available, and it is not yet settled which ones are most suitable.

Background

The prospect of developing A/IS that are sensitive to human norms and factor them into morally or legally significant decisions 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 appears under many names, including: machine morality, machine ethics, moral machines, value alignment, computational ethics, artificial morality, safe AI, and friendly AI.

There are a number of different implementation routes for implementing ethics into autonomous and intelligent systems. Following Wallach and Allen (2008)[14], we might begin to categorize these as either:

A. Top-down approaches, where the system, e.g., a software agent, has some symbolic representation of its activity, and so can identify specific states, plans, or actions as ethical or unethical with respect to particular ethical requirements (Dennis, Fisher, Slavkovik, Webster 2016[15]; Pereira and Saptawijaya 2016[16]; Rötzer, 2016[17]; Scheutz, Malle, and Briggs 2015[18]); or

B. Bottom-up approaches, where the system, e.g., a learning component, builds up, through experience of what is to be considered ethical and unethical in certain situations, an implicit notion of ethical behavior (Anderson and Anderson 2014[19]; Riedl and Harrison 2016[20]).

Relevant examples of these two are: (A) symbolic agents that have explicit representations of plans, actions, goals, etc.; and (B) machine learning systems that train subsymbolic mechanisms with acceptable ethical behavior. For more detailed discussion, see Charisi et al. 2017[21].

Many of the existing experimental approaches to building moral machines are top-down, in the sense that 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. Increasingly, however, A/IS will encounter situations that initially programmed norms do not clearly address, requiring algorithmic procedures to select the better of two or more novel courses of action. Recent breakthroughs in machine learning and perception enable researchers to explore bottom-up approaches in which the A/IS learn about their context and about human norms, similar to the manner in which a child slowly learns which forms of behavior are safe and acceptable. Of course, unlike current A/IS, children can feel pain and pleasure, and empathize with others. Still, A/IS can learn to detect and take into account others' pain and pleasure, thus at least achieving some of the positive effects of empathy. As research on A/IS progresses, engineers will explore new ways to improve these capabilities.

Each of the first two options has obvious limitations, such as option A's inability to learn and adapt and option B's unconstrained learning behavior. A third option tries to address these limitations:

C. Hybrid approaches, combining (A) and (B).

For example, the selection of action might be carried out by a subsymbolic system, but this action must be checked by a symbolic "gateway" agent before being invoked. This is a typical approach for "Ethical Governors" (Arkin, 2008[22]; Winfield, Blum, and Liu 2014[23]) or "Guardians" (Etzioni 2016[24]) that monitor, restrict, and even adapt certain unacceptable behaviors proposed by the system (see Issue 3). Alternatively, action selection in light of norms could be done in a verifiable logical format, while many of the norms constraining those actions can be learned through bottom-up learning mechanisms (Arnold, Kasenberg, and Scheutz 2017[25]).

These three architectures do not cover all possible techniques for implementing norms into A/IS. For example, some contributors to the multi-agent systems literature have integrated norms into their agent specifications (Andrighetto et al. 2013[26]), and even though these agents live in societal simulations and are too underspecified to be translated into individual A/IS such as robots, the emerging work can inform cognitive architectures of such A/IS that fully integrate norms. Of course, none of these experimental systems should be deployed outside of the laboratory before testing or before certain criteria are met, which we outline in the remainder of this section and in Section 3.

Recommendation

In light of the multiple possible approaches to computationally implement norms, diverse research efforts should be pursued, especially collaborative research between scientists from different schools of thought and different disciplines.

Further Resources

- M. Anderson, and S. L. Anderson, "GenEth: A General Ethical Dilemma Analyzer," *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, July 27 –31, 2014, pp. 253–261, Palo Alto, CA, The AAAI Press, 2014.
- G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. Normative Multi-Agent Systems. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.
- R. Arkin, "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, Netherlands, March 12 -15, 2008, IEEE, pp. 121–128, 2008.
- T. Arnold, D. Kasenberg, and M. Scheutz. "Value Alignment or Misalignment—What Will Keep Systems Accountable?" *The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports*, WS-17-02: AI, Ethics, and Society, pp. 81–88. Palo Alto, CA: The AAAI Press, 2017.
- V. Charisi, L. Dennis, M. Fisher, et al. "Towards Moral Autonomous Systems," 2017.
- A. Conn, "How Do We Align Artificial Intelligence with Human Values?" *Future of Life Institute*, Feb. 3, 2017.
- L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal Verification of Ethical Choices in Autonomous Systems." *Robotics and Autonomous Systems,* vol. 77, pp. 1–14, 2016.
- A. Etzioni and O. Etzioni, "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM*, vol. 59, no. 9, pp. 29–31, Sept. 2016.
- L. M. Pereira and A. Saptawijaya, Programming Machine Ethics. Cham, Switzerland: Springer International, 2016.
- M. O. Riedl and B. Harrison. "Using Stories to Teach Human Values to Artificial Agents." *AAAI Workshops 2016*. Phoenix, Arizona, February 12–13, 2016.
- F. Rötzer, ed. Programmierte Ethik: Brauchen Roboter Regeln oder Moral? Hannover, Germany: Heise Medien, 2016.
- M. Scheutz, B. F. Malle, and G. Briggs. "Towards Morally Sensitive Action Selection for Autonomous Social Robots." *Proceedings of the 24th International Symposium on Robot and Human Interactive Communication, RO-MAN 2015* (2015): 492–497.
- U. Sommer, Werte: Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt. [Values. Why we need them even though they don't exist.] Stuttgart, Germany: J. B. Metzler, 2016.
- I. Sommerville, *Software Engineering*. Harlow, U.K.: Pearson Studium, 2001.
- W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.
- F. T. Winfield, C. Blum, and W. Liu. "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection" in *Advances in Autonomous Robotics Systems, Lecture Notes in Computer Science Volume*, M. Mistry, A. Leonardis, Witkowski, and C. Melhuish, eds. pp. 85–96. Springer, 2014.

Issue 2: The need for transparency from implementation to deployment

Background

When A/IS become part of social communities and behave according to the norms of their communities, people will want to understand the A/IS decisions and actions, just as they want to understand each other's decisions and actions. This is particularly true for morally significant actions or omissions: an ethical reasoning system should be able to explain its own reasoning to a user on request. Thus, transparency, or "explainability", of A/IS is paramount (Chaudhuri 2017[27]; Wachter, Mittelstadt, and Floridi 2017[28]), and it will allow a community to understand, predict, and modify the A/IS (see Section 1, Issue 2; for a nuanced discussion see Selbst and Barocas[29]). Moreover, as the norms embedded in A/IS are continuously updated and refined (see Section 1, Issue 2), transparency allows for appropriate trust to be developed (Grodzinsky, Miller, and Wolf 2011[30]), and, where necessary, allows the community to modify a system's norms, reasoning, and behavior.

Transparency can occur at multiple levels, e.g., ordinary language or coder verification, and for multiple stakeholders, e.g., user, engineer, and attorney. (See IEEE P7001™, IEEE Standards Project for Transparency of Autonomous Systems). It should be noted that transparency to all parties may not always be advisable, such as in the case of security programs that prevent a system from being hacked (Kroll et al. 2016[31]). Here we briefly illustrate the broad range of transparency by reference to four ways in which systems can be transparent—traceability, verifiability, honest design, and intelligibility—and apply these considerations to the implementation of norms in A/IS.

Transparency as traceability—Most relevant for the topic of implementation is the transparency of the software engineering process during implementation (Cleland-Huang, Gotel, and Zisman2012[32]). It allows for the originally identified norms (Section 1, Issue 1) to be traced through to the final system. This allows technical inspection of which norms have been implemented, for which contexts, and how norm conflicts are resolved, e.g., priority weights given to different norms. Transparency in the implementation process may also reveal biases that were inadvertently built into systems, such as racism and sexism, in search engine algorithms (Noble 2013[33]). (See Section 3, Issue 2.) Such traceability in turn calibrates a community's trust about whether A/IS are conforming to the norms and values relevant in their use contexts (Fleischmann and Wallace 2005[34]).

Transparency as verifiability—Transparency concerning how normative reasoning is approached in the implementation is important as we wish to verify that the normative decisions the system makes match the required norms and values. Explicit and exact representations of these normative decisions can then provide the basis for a range of strong mathematical techniques, such as formal verification (Fisher, Dennis, and Webster 2013[35]). Even if a system cannot explain every single reasoning step in understandable human terms, a log of ethical reasoning should be available for inspection of later evaluation purposes (Hind et al. 2018[36]).

Transparency as honest design—German designer Dieter Rams coined the term "honest design" to refer to design that "does not make a product more innovative, powerful or valuable than it really is" (Vitsoe 2018[37]; see also Donelli 2015[38]; Jong 2017[39]). Honest design of A/IS is one aspect of their transparency, because it allows the user to "see through" the outward appearance and accurately infer the A/IS' actual capacities. At times, however, the physical appearance of a system does not accurately represent what the system is capable of doing—e.g., the agent displays signs of a certain human-like emotion but its internal state does not represent that human emotion. Humans are quick to make strong inferences from outward appearances of human-likeness to the mental and social capacities the A/IS might have. Demands for transparency in design therefore put a responsibility on the designer to "not attempt to manipulate the consumer with promises that cannot be kept" (Vitsoe 2018[40]).

Transparency as intelligibility—As mentioned above, humans will want to understand the A/IS' decisions and actions, especially the morally significant ones. A clear requirement for an ethical A/IS is that the system be able to explain its own reasoning to a user, when asked—or, ideally, also when suspecting the user's confusion, and the system should do so at a level of ordinary human reasoning, not with incomprehensible technical detail (Tintarev and Kutlak 2014[41]). Furthermore, when the system cannot explain some of its actions, technicians or designers should be available to make those actions intelligible. Along these lines, the European Union's General Data Protection Regulation (GDPR), in effect since May 2018, states that, for 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". (See boyd [sic] 2016[42], for a critical discussion of this regulation.)

Recommendation

A/IS, especially those with embedded norms, must have a high level of transparency, shown as traceability in the implementation process, mathematical verifiability of their reasoning, honesty in appearance-based signals, and intelligibility of the systems' operation and decisions.

Further Resources

- d. boyd, "Transparency ≠ Accountability." *Data & Society: Points*, November 29, 2016.
- A. Chaudhuri, " Philosophical Dimensions of Information and Ethics in the Internet of Things (IoT) Technology,"The EDP Audit, Control, and Security Newsletter, vol. 56, no. 4, pp. 7-18, DOI: 10.1080/07366981.2017.1380474, 2017.
- J. Cleland-Huang, O. Gotel, and A. Zisman, eds. *Software and Systems Traceability.* London: Springer, 2012. doi:10.1007/978-1-4471-2239-5
- G. Donelli, "Good design is honest." (blog). March 13, 2015. Accessed Oct 22, 2018. https://blog.astropad.com/good-design-is-honest/
- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013.

- K. R. Fleischmann and W. A. Wallace. "A Covenant with Transparency: Opening the Black Box of Models." *Communications of the ACM*, vol. 48, no. 5, pp. 93–97, 2005.
- F. S. Grodzinsky, K. W. Miller, and M. J. Wolf. "Developing Artificial Agents Worthy of Trust: Would You Buy a Used Car from This Artificial Agent?" *Ethics and Information Technology*, vol. 13, pp. 17–27, 2011.
- M. Hind, et al. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." *ArXiv E-Prints*, Aug. 2018. [Online] Available: https://arxiv.org/abs/1808.07261. [Accessed October 28, 2018].
- C. W. De Jong, ed., *Dieter Rams: Ten Principles for Good Design*. New York, NY: Prestel Publishing, 2017.
- J. A. Kroll, J. Huey, S. Barocas et al. "Accountable Algorithms." University of Pennsylvania Law Review 165 2017.
- S. U. Noble, "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." InVisible Culture 19, 2013.
- D. Selbst and S. Barocas, "The Intuitive Appeal of Explainable Machines," *87 Fordham Law Review 1085*, Available at SSRN: https://ssrn.com/abstract=3126971 or http://dx.doi.org/10.2139/ssrn.3126971, Feb. 19, 2018.
- N. Tintarev and R. Kutlak. "Demo: Making Plans Scrutable with Argumentation and Natural Language Generation." *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces,* pp. 29–32, 2014.
- Vitsoe. "The Power of Good Design." *Vitsoe*, 2018. Retrieved Oct 22, 2018 from https://www.vitsoe.com/us/about/good-design.
- S.Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accountable AI for Robotics." Science Robotics, vol. 2, no. 6, eaan6080. doi:10.1126/scirobotics. aan6080, 2017.

Issue 3: Failures will occur.

Background

Operational failures and, in particular, violations of a system's embedded community norms, are unavoidable, both during system testing and during deployment. Not only are implementations never perfect, but A/IS with embedded norms will update or expand their norms over time (see Section 1, Issue 2) and interactions in the social world are particularly complex and uncertain. Thus, prevention and mitigation strategies must be adopted, and we sample four possible ones.

First, anticipating the process of evaluation during the implementation phase requires defining criteria and metrics for such evaluation, which in turn better allows the detection and mitigation of failures. Metrics will include:

- Technical variables, such as traceability and verifiability,
- User-level variables such as reliability, understandable explanations, and responsiveness to feedback, and
- Community-level variables such as justified trust (see Issue 2) and the collective belief that A/IS are generally creating social benefits rather than, for example, technological unemployment.

Second, a systematic risk analysis and management approach can be useful (Oetzel and Spiekermann 2014[43]) for an application to privacy norms. This approach tries to anticipate potential points of failure, e.g., norm violations, and, where possible, develops some ways to reduce or remove the effects of failures. Successful behavior, and occasional failures, can then iteratively improve predictions and mitigation attempts.

Third, because not all risks and failures are predictable (Brundage et al 2018[44]; Vanderelst and Winfield 2018[45]), especially in complex human-machine interactions in social contexts, additional mitigation mechanisms must be made available. Designers are strongly encouraged to augment the architectures of their systems with components that handle unanticipated norm violations with a fail-safe, such as the symbolic "gateway" agents discussed in Section 2, Issue 1. Designers should identify a number of strict laws, that is, task- and community-specific norms that should never be violated, and the fail-safe components should continuously monitor operations against possible violations of these laws. In case of violations, the higher-order gateway agent should take appropriate actions, such as safely disabling the system's operation, or greatly limiting its scope of operation, until the source of failure is identified. The fail-safe components need to be understandable, extremely reliable, and protected against security breaches, which can be achieved, for example, by validating them carefully and not letting them adapt their parameters during execution.

Fourth, once failures have occurred, responsible entities, e.g., corporate, government, science, and engineering, shall create a publicly accessible database with undesired outcomes caused by specific A/IS systems. The database would include descriptions of the problem, background information on how the problem was detected, which context it occurred in, and how it was addressed.

In summary, we offer the following recommendation.

Recommendation

Because designers and developers cannot anticipate all possible operating conditions and potential failures of A/IS, multiple strategies to mitigate the chance and magnitude of harm must be in place.

Further Resources

- M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfunkel, A. Dafoe, P. Scharre, T. Zeitzo, et al. " “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” CoRR abs/1802.07228 [cs.AI]. 2018. https://arxiv.org/abs/1802.07228
- M. C. Oetzel and S. Spiekermann, “A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach.” *European Journal of Information Systems*, vol. 23, pp. 126–150, 2014. https://link.springer.com/article/10.1057/ejis.2013.18
- D. Vanderelst and A.F. Winfield, 2018 “The Dark Side of Ethical Robots,” In Proc. The First AAAI/ACM Conf. on Artificial Intelligence, Ethics and Society, New Orleans, LA, Feb. 1 -3, 2018.

Section 3—Evaluating the Implementation of A/IS

The success of implementing appropriate norms in A/IS must be rigorously evaluated. This evaluation process must be anticipated during design and incorporated into the implementation process and continue throughout the life cycle of the system's deployment. Assessment before full-scale deployment would best take place in systematic test beds that allow human users—from the defined community and representing all demographic groups—to engage safely with the A/IS in intended tasks. Multiple disciplines and methods should contribute to developing and conducting such evaluations.

Evaluation criteria must capture, among others, the quality of human-machine interactions, human approval and appreciation of the A/IS, appropriate trust in the A/IS, adaptability of the A/IS to human users, and benefits to human well-being in the presence or under the influence of the A/IS. A range of normative aspects to be considered can be found in British Standard BS 8611:2016 on Robot Ethics (British Standards Institution 2016[46]). These are important general evaluation criteria, but they do not yet fully capture evaluation of a system that has "norm capacities".

To evaluate a system's norm-conforming behavior, one must describe—and ideally, formally specify—criterion behaviors that reflect the previously identified norms, describe what the user expects the system to do, verify that the system really does this, and validate that the specification actually matches the criteria. Many different evaluation techniques are available in the field of software engineering (Sommerville 2015[47]), ranging from formal mathematical proof, through rigorous empirical testing against criteria of normatively correct behavior, to informal analysis of user interactions and responses to the machine's norm awareness and compliance. All these approaches can, in principle, be applied to the full range of A/IS including robots (Fisher, Dennis, and Webster 2013[48]). More general principles from system quality management may also be integrated into the evaluation process, such as the Plan-Do-Check-Act (PDCA) cycle that underlies standards like ISO 9001 (International Organization for Standardization 2015[49]).

Evaluation may be done by first parties, e.g., designers, manufacturers, and users, as well as third parties, e.g., regulators, independent testing agencies, and certification bodies. In either case, the results of evaluations should be made available to all parties, with strong encouragement to resolve discovered system limitations and resolve potential discrepancies among multiple evaluations.

As a general guideline, we recommend that evaluation of A/IS implementations must be anticipated during a system's design, incorporated into the implementation process, and continue throughout the system's deployment (cf. ITIL principles, BMC 2016[50]). Evaluation must include multiple methods, be made available to all parties—from designers and users to regulators, and should include procedures to resolve conflicting evaluation results. Specific issues that need to be addressed in this process are discussed next.

Further Resources

- British Standards Institution. BS8611:2016, "Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems," 2016.
- BMC Software. *ITIL: The Beginner's Guide to Processes & Best Practices*. http://www.bmc.com/guides/itil-introduction.html, Dec. 6, 2016.
- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013.
- International Organization for Standardization (2015). ISO 9001:2015, Quality management systems —Requirements. Retrieved July 12, 2018 from https://www.iso.org/standard/62085.html.
- I. Sommerville, *Software Engineering.* 10th ed. Harlow, U.K.: Pearson Studium, 2015.

Issue 1: Not all norms of a target community apply equally to human and artificial agents

Background

An intuitive criterion for evaluations of norms embedded in A/IS would be that the A/IS norms should mirror the community's norms—that is, the A/IS should be disposed to behave the same way that people expect each other to behave. However, for a given community and a given A/IS use context, A/IS and humans are unlikely to have identical sets of norms. People will have some unique expectations for humans than they do not for machines, e.g., norms governing the regulation of negative emotions, assuming that machines do not have such emotions. People may in some cases have unique expectations of A/IS that they do not have for humans, e.g., a robot worker, but not a human worker, is expected to work without regular breaks.

Recommendation

The norm identification process must document the similarities and differences between the norms that humans apply to other humans and the norms they apply to A/IS. Norm implementations should be evaluated specifically against the norms that the community expects the A/IS to follow.

Issue 2: A/IS can have biases that disadvantage specific groups

Background

Even when reflecting the full system of community norms that was identified, A/IS may show operation biases that disadvantage specific groups in the community or instill biases in users by reinforcing group stereotypes. A system's bias can emerge in perception. For example, a passport application AI rejected an Asian man's photo because it insisted his eyes were closed (Griffiths 2016[51]). Bias can emerge in information processing. For instance, speech recognition systems are notoriously less accurate for female speakers than for male speakers (Tatman 2016[52]). System bias can affect decisions, such as a criminal risk assessment device which overpredicts recidivism by African Americans (Angwin et al. 2016[53]). The system's bias can present itself even in its own appearance and presentation: the vast majority of humanoid robots have white "skin" color and use female voices (Riek and Howard 2014[54]).

The norm identification process detailed in Section 1 is intended to minimize individual designers' biases because the community norms are assessed empirically. The identification process also seeks to incorporate norms against prejudice and discrimination. However, biases may still emerge from imperfections in the norm identification process itself, from unrepresentative training sets for machine learning systems, and from programmers' and designers' unconscious assumptions. Therefore, unanticipated or undetected biases should be further reduced by including members of diverse social groups in both the planning and evaluation of A/IS and integrating community outreach into the evaluation process, e.g., DO-IT program and RRI framework. Behavioral scientists and members of the target populations will be particularly valuable when devising criterion tasks for system evaluation and assessing the success of evaluating the A/IS performance on those tasks. Such tasks would assess, for example, whether the A/IS apply norms in discriminatory ways to different races, ethnicities, genders, ages, body shapes, or to people who use wheelchairs or prosthetics, and so on.

Recommendation

Evaluation of A/IS must carefully assess potential biases in the systems' performance that disadvantage specific social and demographic groups. The evaluation process should integrate members of potentially disadvantaged groups in efforts to diagnose and correct such biases.

Further Resources

- J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." ProPublica, May 23, 2016.
- J. Griffiths, "New Zealand Passport Robot Thinks This Asian Man's Eyes Are Closed." CNN.com, December 9, 2016.

- L. D. Riek and D. Howard,. “A Code of Ethics for the Human-Robot Interaction Profession.” *Proceedings of We Robot,* April 4, 2014.
- R. Tatman, “Google’s Speech Recognition Has a Gender Bias.” *Making Noise and Hearing Things*, July 12, 2016.

Issue 3: Challenges to evaluation by third parties

Background

A/IS should have sufficient transparency to allow evaluation by third parties, including regulators, consumer advocates, ethicists, post-accident investigators, or society at large. However, transparency can be severely limited in some systems, especially in those that rely on machine learning algorithms trained on large data sets. The data sets may not be accessible to evaluators; the algorithms may be proprietary information or mathematically so complex that they defy common-sense explanation; and even fellow software experts may be unable to verify reliability and efficacy of the final system because the system’s specifications are opaque.

For less inscrutable systems, numerous techniques are available to evaluate the implementation of the A/IS’ norm conformity. On one side there is formal verification, which provides a mathematical proof that the A/IS will always match specific normative and ethical requirements, typically devised in a top-down approach (see Section 2, Issue 1). This approach requires access to the decision-making process and the reasons for each decision (Fisher, Dennis, and Webster 2013[55]). A simpler alternative, sometimes suitable even for machine learning systems, is to test the A/IS against a set of scenarios and assess how well they matches their normative requirements, e.g., acting in accordance with relevant norms and recognizing other agents’ norm violations. A “red team” may also devise scenarios that try to get the A/IS to break norms so that its vulnerabilities can be revealed.

These different evaluation techniques can be assigned different levels of “strength”: strong ones demonstrate the exhaustive set of the A/IS’ allowable behaviors for a range of criterion scenarios; weaker ones sample from criterion scenarios and illustrate the systems’ behavior for that subsample. In the latter case, confidence in the A/IS’ ability to meet normative requirements is more limited. An evaluation’s concluding judgment must therefore acknowledge the strength of the verification technique used, and the expressed confidence in the evaluation—and in the A/IS themselves—must be qualified by this level of strength.

Transparency is only a necessary requirement for a more important long-term goal: having systems be accountable to their users and community members. However, this goal raises many questions such as to whom the A/IS are accountable, who has the right to correct the systems, and which kind of A/IS should be subject to accountability requirements.

Recommendation

To maximize effective evaluation by third parties, e.g., regulators and accident investigators, A/IS should be designed, specified, and documented so as to permit the use of strong verification and validation techniques for assessing the system's safety and norm compliance, in order to achieve accountability to the relevant communities.

Further Resources

- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM,* vol. 56, pp. 84–93, 2013.
- K. Abney, G. A. Bekey, and P. Lin. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: The MIT Press, 2011.
- M. Anderson and S. L. Anderson, eds. *Machine Ethics.* New York: Cambridge University Press, 2011.
- M. Boden, J. Bryson, et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science* 29, no. 2, pp. 124–129, 2017.
- M. Coeckelbergh, "Can We Trust Robots?" *Ethics and Information Technology,* vol.14, pp. 53–60, 2012.
- L. A. Dennis, M. Fisher, N. Lincoln, A. Lisitsa, and S. M. Veres, "Practical Verification of Decision-Making in Agent-Based Autonomous Systems." *Automated Software Engineering,* vol. 23, no. 3, pp. 305–359, 2016.
- M. Fisher, C. List, M. Slavkovik, and A. F. T. Winfield. "Engineering Moral Agents—From Human Morality to Artificial Morality" (Dagstuhl Seminar 16222). *Dagstuhl Reports* 6, no. 5, pp. 114–137, 2016.
- K. R. Fleischmann, *Information and Human Values*. San Rafael, CA: Morgan and Claypool, 2014.
- G. Governatori and A. Rotolo. "How Do Agents Comply with Norms? " in *Normative Multi-Agent Systems*, G. Boella, P. Noriega, G. Pigozzi, and H. Verhagen, eds., *Dagstuhl Seminar Proceedings*. Dagstuhl, Germany: Schloss Dagstuhl—Leibniz- Zentrum für Informatik, 2009.
- B. Higgins, "New York City Task Force to Consider Algorithmic Harm." *Artificial Intelligence Technology and the Law Blog*, Feb. 7, 2018. [Online]. Available: http://aitechnologylaw.com/2018/02/new-york-city-task-force-algorithmic-harm/. [Accessed Nov. 1, 2018].
- S. L. Jarvenpaa, N. Tractinsky, and L. Saarinen. "Consumer Trust in an Internet Store: A Cross-Cultural Validation" *Journal of Computer-Mediated Communication,* vol. 5, no. 2, pp. 1–37, 1999.
- E. H. Leet and W. A. Wallace. "Society's Role and the Ethics of Modeling," in *Ethics in Modeling*, W. A. Wallace, ed., Tarrytown, NY: Elsevier, 1994, pp. 242– 245.
- M. A. Mahmoud, M. S. Ahmad, M. Z. M. Yusoff, and A. Mustapha. "A Review of Norms and Normative Multiagent Systems," *The Scientific World Journal*, vol. 2014, Article ID 684587, 2014.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Embedding Values into Autonomous Intelligent Systems Committee

- **AJung Moon** (Founding Chair) – Director of Open Roboethics Institute
- **Bertram F. Malle** (Co-Chair) – Professor, Department of Cognitive, Linguistic, and Psychological Sciences, Co-Director of the Humanity-Centered Robotics Initiative, Brown University
- **Francesca Rossi** (Co-Chair) – Full Professor, computer science at the University of Padova, Italy, currently at the IBM Research Center at Yorktown Heights, NY
- **Stefano Albrecht** – Postdoctoral Fellow in the Department of Computer Science at The University of Texas at Austin
- **Bijilash Babu** – Senior Manager, Ernst and Young, EY Global Delivery Services India LLP
- **Jan Carlo Barca** – Senior Lecturer in Software Engineering and Internet of Things (IoT), School of Info Technology, Deakin University, Australia
- **Catherine Berger** – IEEE Standards Senior Program Manager, IEEE
- **Malo Bourgon** – COO, Machine Intelligence Research Institute
- **Richard S. Bowyer** – Adjunct Senior Lecturer and Research Fellow, College of Science and Engineering, Centre for Maritime Engineering, Control and Imaging (cmeci), Flinders University, South Australia
- **Stephen Cave** – Executive Director of the Leverhulme Centre for the Future of Intelligence, University of Cambridge
- **Raja Chatila** – CNRS-Sorbonne Institute of Intelligent Systems and Robotics, Paris, France; Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA; Past President of IEEE Robotics and Automation Society
- **Mark Coeckelbergh** – Professor, Philosophy of Media and Technology, the University of Vienna
- **Louise Dennis** – Lecturer, Autonomy and Verification Laboratory, University of Liverpool
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions'; member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Virginia Dignum** – Associate Professor, Faculty of Technology Policy and Management, TU Delft

- **Ebru Dogan** – Research Engineer, VEDECOM
- **Takashi Egawa** – Cloud Infrastructure Laboratory, NEC Corporation, Tokyo
- **Vanessa Evers** – Professor, Human-Machine Interaction, and Science Director, DesignLab, University of Twente
- **Michael Fisher** – Professor of Computer Science, University of Liverpool, and Director of the UK Network on the Verification and Validation of Autonomous Systems, vavas.org
- **Ken Fleischmann** – Associate Professor in the School of Information at The University of Texas at Austin
- **Edith Pulido Herrera** – Bioengineering group, Antonio Nariño University, Bogotá, Colombia
- **Ryan Integlia** – assistant professor, Electrical and Computer Engineering, Florida Polytechnic University; Co-Founder of the em[POWER] Energy Group
- **Catholijn Jonker** – Full professor of Interactive Intelligence at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology. Part-time full professor at Leiden Institute of Advanced Computer Science of the Leiden University
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech
- **Jong-Wook Kim** – Professor, AI.Robotics Lab, Department of Electronic Engineering, Dong-A University, Busan, Korea
- **Sven Koenig** – Professor, Computer Science Department, University of Southern California
- **Brenda Leong** – Senior Counsel, Director of Operations, The Future of Privacy Forum
- **Alan Mackworth** – Professor of Computer Science, University of British Columbia; Former President, AAAI; Co-author of “Artificial Intelligence: Foundations of Computational Agents”.
- **Pablo Noriega** – Scientist, Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC), Barcelona.
- **Rajendran Parthiban** – Professor, School of Engineering, Monash University, Bandar Sunway, Malaysia
- **Heather M. Patterson** – Senior Research Scientist, Anticipatory Computing Lab, Intel Corp.
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow.
- **Laurel Riek** – Associate Professor, Computer Science and Engineering, University of California San Diego
- **Leanne Seeto** – Co-Founder and Strategy and Operations Precision Autonomy
- **Sarah Spiekermann** – Chair of the Institute for Information Systems & Society at Vienna University of Economics and Business; Author of the textbook “Ethical IT-Innovation”, the popular book “Digitale Ethik—Ein Wertesystem für das 21. Jahrhundert” and Blogger on ”The Ethical Machine”

- **John P. Sullins** – Professor of Philosophy, Chair of the Center for Ethics Law and Society (CELS), Sonoma State University
- **Jaan Tallinn** – Founding engineer of Skype and Kazaa; co-founder of the Future of Life Institute
- **Mike Van der Loos** – Associate Prof., Dept. of Mechanical Engineering, Director of Robotics for Rehabilitation, Exercise and Assessment in Collaborative Healthcare (RREACH) Lab, and Associate Director of CARIS Lab, University of British Columbia
- **Wendell Wallach** – Consultant, ethicist, and scholar, Yale University's Interdisciplinary Center for Bioethics
- **Nell Watson** – CFBCS, FICS, FIAP, FIKE, FRSA, FRSS, FLS Co-Founder and Chairman, EthicsNet, AI & Robotics Faculty Singularity University, Foresight Machine Ethics Fellow
- **Karolina Zawieska** – Postdoctoral Research Fellow in Ethics and Cultural Learning of Robotics at DeMontfort University, UK and Researcher at Industrial Research Institute for Automation and Measurements PIAP, Poland

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see *How the Document Was Prepared*.

Endnotes

[1] S. Hitlin and J. A. Piliavin. “Values: Reviving a Dormant Concept.” *Annual Review of Sociology* 30 (2004): 359–393.

[2] B. F. Malle, and S. Dickert. “Values,” *The Encyclopedia of Social Psychology*, edited by R. F. Baumeister and K. D. Vohs. Thousand Oaks, CA: Sage, 2007.

[3] M. J. Rohan, “A Rose by Any Name? The Values Construct.” *Personality and Social Psychology Review* 4 (2000): 255–277.

[4] A. U. Sommer, Werte: *Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt.* [Values. Why We Need Them Even Though They Don’t Exist.] Stuttgart, Germany: J. B. Metzler, 2016.

[5] B. F. Malle, M. Scheutz, and J. L. Austerweil. “Networks of Social and Moral Norms in Human and Robot Agents,” in A World with Robots: *International Conference on Robot Ethics*: ICRE 2015, edited by M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, and G. S. Virk, 3–17. Cham, Switzerland: Springer International Publishing, 2017.

[6] J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. “Implementing Norms in Multiagent Systems,” in Multiagent System Technologies. MATES 2004, edited by G. Lindemann, Denzinger, I. J. Timm, and R. Unland. (Lecture Notes in Computer Science, vol. 3187.) Berlin: Springer, 2004.

[7] A. Mack, (Ed.). “Changing social norms.” *Social Research: An International Quarterly,* 85, no.1 (2018): 1–271.

[8] I. van de Poel, “An Ethical Framework for Evaluating Experimental Technology”, *Science and Engineering Ethics*, 22, no. 3 (2016): 667-686.

[9] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, (2016). Seeing through the human reporting bias: Visual Classifiers from Noisy Human-Centric Labels. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 2930–2939). doi:10.1109/CVPR.2016.320

[10] A. Mack, (Ed.). (2018). Changing social norms. *Social Research: An International Quarterly*, 85(1, Special Issue), 1–271.

[11] B. Green and L. Hu. “The Myth in the Methodology: Towards a Recontextualization of Fairness in ML.” Paper presented at the Debates workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden 2018.

[12] J. Van den Hoven, “Engineering and the Problem of Moral Overload.” *Science and Engineering Ethics* 18, no. 1 (2012): 143–155.

[13] C. Andre and M. Velasquez. “The Common Good.” *Issues in Ethics* 5, no. 1 (1992).

[14] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press, 2008.

[15] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. “Formal Verification of Ethical Choices in Autonomous Systems.” *Robotics and Autonomous Systems* 77 (2016): 1–14.

[16] L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics*. Cham, Switzerland: Springer International, 2016.

[17] F. Rötzer, ed. Programmierte Ethik: Brauchen Roboter Regeln oder Moral? Hannover, Germany: Heise Medien, 2016.

[18] M. Scheutz, B. F. Malle, and G. Briggs. "Towards Morally Sensitive Action Selection for Autonomous Social Robots." *Proceedings of the 24th International Symposium on Robot and Human Interactive Communication, RO-MAN 2015* (2015): 492–497.

[19] M. Anderson and S. L. Anderson. "GenEth: A General Ethical Dilemma Analyzer." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014): 253–261.

[20] M. O. Riedl and B. Harrison. "Using Stories to Teach Human Values to Artificial Agents." *Proceedings of the 2nd International Workshop on AI, Ethics and Society*, Phoenix, Arizona, 2016.

[21] V. Charisi, L. Dennis, M. Fisher et al. "Towards Moral Autonomous Systems," 2017.

[22] R. Arkin, "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction* (2008): 121–128.

[23] A. F. T. Winfield, C. Blum, and W. Liu. "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection" in *Advances in Autonomous Robotics Systems, Lecture Notes in Computer Science Volume*, edited by M. Mistry, A. Leonardis, Witkowski, and C. Melhuish, 85–96. Springer, 2014.

[24] A. Etzioni, "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM* 59, no. 9 (2016): 29–31.

[25] T. Arnold, D. Kasenberg, and M. Scheutz. "Value Alignment or Misalignment—What Will Keep Systems Accountable?" The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports, WS-17-02: AI, Ethics, and Society, 81–88. Palo Alto, CA: The AAAI Press, 2017.

[26] G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. *Normative Multi-Agent Systems*. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.

[27] A. Chaudhuri, (2017) Philosophical Dimensions of Information and Ethics in the Internet of Things (IoT) Technology. The EDP Audit, Control, and Security Newsletter, 56:4, 7-18, DOI: 10.1080/07366981.2017.1380474

[28] S.Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accountable AI for Robotics." Science Robotics 2, no. 6 (2017): eaan6080. doi:10.1126/scirobotics. aan6080

[29] A. D. Selbst and S. Barocas, The Intuitive Appeal of Explainable Machines (February 19, 2018). Fordham Law Review. Available at SSRN: https://ssrn.com/abstract=3126971 or http://dx.doi.org/10.2139/ssrn.3126971

[30] F. S. Grodzinsky, K. W. Miller, and M. J. Wolf. "Developing Artificial Agents Worthy of Trust: Would You Buy a Used Car from This Artificial Agent?" Ethics and Information Technology 13, (2011): 17–27.

[31] J. A. Kroll, J. Huey, S. Barocas et al. “Accountable Algorithms.” University of Pennsylvania Law Review 165 (2017).

[32] J. Cleland-Huang, O. Gotel, and A. Zisman, eds. Software and Systems Traceability. London: Springer, 2012. doi:10.1007/978- 1-4471-2239-5

[33] S. U. Noble, “Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible.” InVisible Culture 19 (2013).

[34] K. R. Fleischmann and W. A. Wallace. “A Covenant with Transparency: Opening the Black Box of Models.” Communications of the ACM 48, no. 5 (2005): 93–97.

[35] M. Fisher, L. A. Dennis, and M. P. Webster. “Verifying Autonomous Systems.” Communications of the ACM 56, no. 9 (2013): 84–93.

[36] M. Hind, et al. “Increasing Trust in AI Services through Supplier’s Declarations of Conformity.” *ArXiv E-Prints*, Aug. 2018. Retrieved October 28, 2018 from https://arxiv.org/abs/1808.07261.

[37] Vitsoe. “The Power of Good Design.” *Vitsoe*, 2018. Retrieved Oct 22, 2018 from https://www.vitsoe.com/us/about/good-design.

[38] G. Donelli, (2015, March 13). Good design is honest (Blogpost). Retrieved Oct 22, 2018 from https://blog.astropad.com/good-design-is-honest/

[39] C. de Jong Ed., “Ten principles for good design: Dieter Rams.” New York, NY: Prestel Publishing, 2017.

[40] Ibid.

[41] N. Tintarev and R. Kutlak. “Demo: Making Plans Scrutable with Argumentation and Natural Language Generation.” *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces* (2014): 29–32.

[42] d. boyd, “Transparency ≠ Accountability.” *Data & Society: Points*, November 29, 2016.

[43] C. Oetzel and S. Spiekermann, “A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach.” *European Journal of Information Systems* 23, (2014): 126–150. https://link.springer.com/article/10.1057/ejis.2013.18

[44] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfunkel, A. Dafoe, P. Scharre, T. Zeitzo, et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. CoRR abs/1802.07228 (2018). https://arxiv.org/abs/1802.07228M.

[45] D. Vanderelst and A.F. Winfield, 2018 The Dark Side of Ethical Robots. In Proc. AAAI/ACM Conf. on Artificial Intelligence, Ethics and Society, New Orleans.

[46] British Standards Institution. BS8611:2016, “Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems,” 2016.

[47] I. Sommerville, *Software Engineering* (10th edition). Harlow, U.K.: Pearson Studium, 2015.

[48] M. Fisher, L. A. Dennis, and M. P. Webster. “Verifying Autonomous Systems.” Communications of the ACM 56, no. 9 (2013): 84–93.

[49] International Organization for Standardization (2015). ISO 9001:2015, Quality management systems—Requirements. Retrieved July 12, 2018 from https://www.iso.org/standard/62085.html.

[50] BMC Software. *ITIL: The Beginner's Guide to Processes & Best Practices*. 6 Dec. 2016, http://www.bmc.com/guides/itil-introduction.html.

[51] J. Griffiths, "New Zealand Passport Robot Thinks This Asian Man's Eyes Are -Closed." CNN.com, December 9, 2016.

[52] R. Tatman, "Google's Speech Recognition Has a Gender Bias." Making Noise and Hearing Things, July 12, 2016.

[53] J. Angwin, J. Larson, S. Mattu, L. Kirchner. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." ProPublica, May 23, 2016.

[54] L. D. Riek and D. Howard. "A Code of Ethics for the Human-Robot Interaction Profession." Proceedings of We Robot, April 4, 2014.

[55] M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." Communications of the ACM 56 (2013): 84–93.