Embedding Values into Autonomous Intelligent Systems

Society has not established universal standards or guidelines for embedding human norms and values 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 they be designed to adopt, learn, and follow the norms and values of the community they serve. Moreover, their actions must 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 levels of trust in A/IS that are appropriate in the specific contexts and roles in which A/IS function.

The conceptual complexities surrounding what “values” are (e.g., Hitlin and Piliavin, 2004; Malle and Dickert, 2007; Rohan, 2000; Sommer, 2016) make it currently difficult to envision A/IS that have computational structures directly corresponding to social or cultural values (such as “security,” “autonomy,” or “fairness”). However, it is a more realistic goal to embed explicit norms into such systems because norms can be considered 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 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, even if there are no directly identifiable computational structures that correspond to values per se. (For discussion of specific values that are critical for ethical considerations of A/IS, see the sections “Personal Data and Individual Access Control” and “Well-being”.)

Norms are typically expressed in terms of obligations and prohibitions, and these can be expressed computationally (e.g., Malle, Scheutz, and Austerweil, 2017; Vázquez-Salceda, Aldewereld, Dignum, 2004). At this level, norms 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.

Embedding Values into Autonomous Intelligent Systems

To address the broad objective of embedding norms and, by implication, values into these systems, our Committee has defined three more concrete goals as described in the following sections:

1. Identifying the norms of a specific community in which A/IS operate.

2. Computationally implementing the norms of that community within the A/IS.

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 A/IS and their users within a specific community. It is understood that there may be clashes of values and norms when identifying, implementing, and evaluating these systems. Such clashes are a natural part of the dynamically changing and renegotiated norm systems of any community. As a result, we advocate for an approach where systems are designed to provide transparent signals (such as explanations or inspection capabilities) about the specific nature of their behavior to the individuals in the community they serve.

References

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

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

- Malle, B. F., 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.

Embedding Values into Autonomous Intelligent Systems

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

- Sommer, A. U. *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.

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

Disclaimer: *While we have provided recommendations in this document, it should be understood these do not represent a position or the views of IEEE but the informed opinions of Committee members providing insights designed to provide expert directional guidance regarding A/IS. In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors or omissions, direct or otherwise, however caused, arising in any way out of the use of this work, regardless of whether such damage was foreseeable.*

Section 1 — Identifying Norms for Autonomous Intelligent Systems

We identify three issues that must be addressed in the attempt to identify norms (and thereby 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 fundamental property of norms: their dynamically changing nature, 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.

Issue 1: Which norms should be identified?

Background and Analysis

If machines engage in 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. But which norms? Laws are publicly documented and therefore easy to identify, so they will certainly have to be incorporated into A/IS. Social and moral norms are more difficult to ascertain, as they are expressed through behavior, language, customs, cultural symbols, and artifacts. Most important, communities (from families to whole nations) differ to various degrees in the laws and norms they follow. Therefore, generating a universal set of norms that applies to all autonomous systems is not realistic, but neither is it advisable to completely personalize an A/IS to individual preferences. However, we believe that identifying broadly observed norms of a particular community is feasible.

The difficulty of generating a set of universal norms is not inconsistent with the goal of seeking agreement over Universal Human Rights (see "General Principles" section). However, such universal rights would not be sufficient for devising an A/IS that obeys the specific norms of its community. Universal rights must, however, constrain the kinds of norms that are implemented in an A/IS.

Embedding norms in A/IS requires a clear delineation of the community 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, will differ greatly from those for robots used in healthcare. Thus, we recommend that 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 that the A/IS are designed to perform. Even when designating a narrowly defined community (e.g., a nursing home; an apartment complex; a company), there will be variations in the norms that apply. The 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 norms of a given individual, such as norms of arranging a disabled person's home to accommodate certain physical limitations. However, unique individual norms 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.)

We should note that the norms that apply to humans may not always be identical to the norms that would apply to an A/IS in the same context. Empirical research involving multiple disciplines and multiple methods (see the Further Resources section) should therefore (a) investigate and document both community- and task-specific norms that apply to humans and (b) consider possible differences for A/IS deployed in these contexts. The set of empirically identified norms applicable to A/IS should then be made available for designers to implement.

Candidate 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 that the A/IS are designed to perform.

Further Resources

- Bendel, O. *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.

- Burks, S. V., 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* 58 (2012): 203–217. Illustrates surveys and incentivized coordination games as methods to elicit norms in a large financial services firm.

- Friedman, B., P. H. Kahn, A. Borning, and A. Huldtgren. "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory* (Vol. 16), edited by N. Doorn, D. Schuurbiers, I. van de Poel, and M. E. Gorman, 55–95. Dordrecht: Springer, 2013. A comprehensive introduction into Value Sensitive Design and three sample applications.

- Mackie, G., 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, 2012. A broad survey of conceptual and measurement questions regarding social norms.

- Malle, B. F. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18, no. 4 (2016): 243–256. Discusses how a robot's norm capacity fits in the larger vision of a robot with moral competence.

- Miller, K. W., 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* 23 (2017): 389–401. This article raises doubts about the possibility of imbuing artificial agents with morality, or claiming to have done so.

- Rizzo, A., 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* 14 (2004): 335–348. Describes two assessment instruments of moral reasoning (including norm maintenance) based on Kohlberg's theory of moral development.

- Schwartz, S. H. "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).

- Schwartz, S. H., and K. Boehnke. "Evaluating the Structure of Human Values with Confirmatory Factor Analysis." *Journal of Research in Personality 38* (2004): 230–255. Describes an older method of subjective judgments of relations among valued outcomes and a newer, formal method of analyzing these relations.

- Wallach, W., 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.

Issue 2:
The need for norm updating.

Background and Analysis

Norms are not static. They change over time, in response to social progress and new legal measures, and, in smaller communities, in response to complaints or new opportunities. New norms form when technological innovation demands novel social standards (e.g., cell phone use in public), and norms can fade away when, for whatever reasons, fewer and fewer people adhere to them.

Humans have many mechanisms available to update 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, a job in a new town); and they respond to feedback from others when they exhibit uncertainty about norms or have violated a norm.

An A/IS may be equipped with a norm baseline before it is deployed in its target community (Issue 1), but this will not suffice for it to behave appropriately over an extended time. It must be capable of identifying and adding new norms to its baseline system, because the initial norm identification process will undoubtedly have missed some norms. It must also be capable of updating some of its existing norms, as change occurs in its target community. A/IS would be best equipped to respond to such demands for change by relying on multiple mechanisms, such as:

- 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 the robot to a previously unknown context or who notice the A/IS's uncertainty in a familiar context;
- Responding to critique from the community when the A/IS violates 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.

As in the case of resolving norm conflicts (Issue 2), we recommend that the system's norm changes be transparent. That is, the system should make explicit when it adds new norms to its norm system or adjusts the priority or content of existing norms. The specific form of communication will vary by machine sophistication (e.g., communication capacity) and function (e.g., flexible social companion vs. task-defined medical robot). 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 may be appropriate; in yet other cases, the A/IS may propose changes and the relevant human community will decide whether such changes should be implemented in the system.

Candidate Recommendation

To respond to the dynamic change of norms in society the A/IS must be able to adjust its existing norms and learn new ones, while being transparent about these changes.

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

Background and Analysis

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; see Van den Hoven, 2012) must be computationally tractable by an A/IS — it cannot simply stop in its 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, we need to 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. Some general principles are available, such as the Common Good Principle (Andre and Velasquez, 1992). However, other priority relations in the norm network must be established through empirical research so as to reflect the shared values of the community in question. 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.

Some priority orders can be built into a given norm network as hierarchical relations (e.g., prohibitions against harm to humans typically override prohibitions against lying). Other priority orders can stem from the general 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), an A/IS of a racist user who demands the A/IS use derogatory language for certain social groups might 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 input (e.g., about the subtle differences between contexts) to refine the system's norm network (see Issue 2).

We also 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. Just like people explain to each other why they made decisions, they will expect any A/IS to be able to explain its decisions (and be sensitive to user feedback about the appropriateness of the decision). To do so, design and development of A/IS should specifically identify the relevant groups of humans who may request explanations and evaluate the system's behavior.

Candidate Recommendation

One must identify the ways in which people resolve norm conflicts and the ways in which they expect A/IS to resolve similar norm conflicts. The system's resolution of norm conflicts must be transparent — that is, documented by the system and ready to be made available to relevant users.

References

- Andre, C., and M. Velasquez. "The Common Good." *Issues in Ethics* 5, no. 1 (1991).
- Van den Hoven, J. "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics* 18, no. 1 (2012): 143–155.

Further Resources

- Abel, D., 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.
- Cushman, F., V. Kumar, and P. Railton. "Moral Learning." *Cognition* 167 (2017): 1–282.
- Open Roboethics Initiative (e.g., on care robots). A series of poll results on differences in human moral decision-making and changes in priority order of values for autonomous systems.

Section 2 — Implementing Norms in Autonomous Intelligent Systems

Once the norms relevant to an A/IS's 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. Before we discuss these three issues and corresponding candidate recommendations, we offer one general recommendation for the entire process of implementation:

Candidate Recommendation

Throughout the technical implementation of norms, designers should already consider forms and metrics of evaluation and define and incorporate central criteria for assessing an A/IS's norm conformity (e.g., human-machine agreement on moral decisions, verifiability of A/IS decisions, justified trust).

Issue 1:

Many approaches to norm implementation are currently available, and new ones are being developed.

Background and Analysis

The prospect of developing artificial systems 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 systems. Following Wallach and Allen (2008), 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/unethical with respect to particular ethical requirements (e.g., Dennis, Fisher, Slavkovik, Webster, 2016; Pereira and Saptawijaya, 2016; Rötzer, 2016; Scheutz, Malle, and Briggs, 2015); or

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

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.)

Computers and robots already reflect values in their choices and actions, but these values are programmed or designed in 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. 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.

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 norms, 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 an AI system cannot presently imitate. Nevertheless, as research on autonomous systems progresses, engineers will explore new ways to either simulate learning capabilities or build alternative mechanisms that fulfill similar functions.

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; Winfield, Blum, and Liu, 2014) or Guardians (Etzioni, 2016) that monitor, restrict, and even adapt certain unacceptable behaviors proposed by the system. (See also 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 (e.g., Arnold, Kasenberg, and Scheutz, 2017).

These three architectures are not a comprehensive list of all possible techniques for implementing norms and values into A/IS. For example, some contributors to the multi-agent systems literature have integrated norms into their agent specifications (Andrighetto et al., 2013), 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. In addition, some experimental approaches may attempt to capture values computationally (Conn, 2017), or attempt to relate norms to values in ways that ground or justify norms (Sommer, 2016). 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.

Candidate 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.

Issue 2:
The need for transparency from implementation to deployment.

Background and Analysis

When A/IS are part of social communities and act 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 (Wachter, Mittelstadt, and Floridi, 2017), and it will allow a community to understand, predict, and appropriately trust the A/IS (see Section 1, Issue 2). Moreover, as the norms embedded in A/IS are continuously updated and refined (see Section 1, Issue 2), transparency allows for trust to be maintained (Grodzinsky, Miller, and Wolf 2011), 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, coder verification) and for multiple stakeholders (e.g., user, engineer, attorney). (See IEEE P7001™, Draft Standard 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). Here we briefly illustrate the broad range of transparency by reference to four ways in which systems can be transparent (traceability, verifiability, nondeception, 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 Zisman, 2012). 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 (e.g., Noble, 2013). (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 its use context (Fleischmann and Wallace, 2005).

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). 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.

Transparency as nondeception and honest design. We can assume that lying and deception will be prohibited actions in many contexts, and therefore will be part of the norm system implemented into A/IS. In certain use cases of an A/IS, deception may be necessary in serving the core functionality of the system (e.g., a robot that plays poker with humans), but those actions are no longer norm violations because they are justified by context and user consent.

However, the absence of deception does not yet meet the goal of transparency. One should demand that A/IS be *honest*, and that includes both, more obviously, honest communication by the A/IS itself and, less obviously, "honest design." Honest design entails that the physical appearance of a system accurately represents what the system is capable of doing — e.g., ears only for systems that actually process acoustic information; eyes only for systems that actually process visual information. The requirement for honest design may also extend to higher-level capacities of artificial agents: If the agent introduces a certain topic into conversation, then it should also be able to, if asked, reason about that topic; if the agent displays signs of a certain human-like emotion, then it should have an internal state that corresponds to at least an analogue to that human emotion (e.g., inhabit the appraisal states that make up the emotion).

Transparency as intelligibility. As mentioned above, humans will want to understand an A/IS's decisions and actions, especially the morally significant ones. A clear requirement for an ethical A/IS is therefore 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). Furthermore, when the system cannot itself explain some of its actions, technicians or designers should be available to make those actions intelligible. Along these lines, the European Union's new General Data Protection Regulation (GDPR), scheduled to take effect in 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, 2016, for a critical discussion of this regulation.)

Candidate Recommendation

A/IS, and especially those with embedded norms, must have a high level of transparency, from traceability in the implementation process, mathematical verifiability of its reasoning, to honesty in appearance-based signals, and intelligibility of the system's operation and decisions.

Issue 3:
Failures will occur.

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 extended use (see Section 1, Issue 2) and interactions in the social world are particularly complex and uncertain. Thus, we propose the following candidate recommendation.

Candidate Recommendation

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

Elaboration

To be specific, we sample three possible mitigation strategies.

First, anticipating the process of evaluation already 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 more 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 (e.g., Oetzel and Spiekermann, 2014, 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 mitigate 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, 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 1, 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 until the source of failure is identified. The fail-safe components need to be 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.

References

- Anderson, M., and S. L. Anderson. "GenEth: A General Ethical Dilemma Analyzer." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014): 253–261.
- Andrighetto, G., G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. *Normative Multi-Agent Systems*. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.
- Arkin, R. "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." *Proceedings of the 2008 Conference on Human-Robot Interaction* (2008): 121–128.
- Arnold, T., 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.
- Boyd, D. "Transparency ≠ Accountability." *Data & Society: Points*, November 29, 2016.
- Charisi, V., L. Dennis, M. Fisher et al. "Towards Moral Autonomous Systems," 2017.
- Cleland-Huang, J., O. Gotel, and A. Zisman, eds. *Software and Systems Traceability*. London: Springer, 2012. doi:10.1007/978-1-4471-2239-5
- Conn, A. "How Do We Align Artificial Intelligence with Human Values?" *Future of Life Institute*, February 3, 2017.

- Dennis, L., M. Fisher, M. Slavkovik, and M. Webster. "Formal Verification of Ethical Choices in Autonomous Systems." *Robotics and Autonomous Systems* 77 (2016): 1–14.
- Etzioni, A. "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM* 59, no. 9 (2016): 29–31.
- Fisher, M., L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56, no. 9 (2013): 84–93.
- Fleischmann, K. R., and W. A. Wallace. "A Covenant with Transparency: Opening the Black Box of Models." *Communications of the ACM* 48, no. 5 (2005): 93–97.
- Grodzinsky, F. S., 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.
- Kroll, J. A., J. Huey, J., S. Barocas et al. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (2017 forthcoming).
- Noble, S. U. "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." *InVisible Culture* 19 (2013).
- Oetzel, M. C., and S. Spiekermann. "A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach." *European Journal of Information Systems* 23, (2014): 126–150. doi:10.1057/ejis.2013.18
- Pereira, L. M., and A. Saptawijaya. *Programming Machine Ethics*. Cham, Switzerland: Springer International, 2016.
- Riedl, M. O., 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.
- Rötzer, F. ed. *Programmierte Ethik: Brauchen Roboter Regeln oder Moral?* Hannover, Germany: Heise Medien, 2016.
- Scheutz, M., 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.
- Sommer, A. U. *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.
- Sommerville, I. *Software Engineering*. Harlow, U.K.: Pearson Studium, 2001.
- Tintarev, N., 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.

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

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

- Winfield A. F. T., 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, M. Witkowski, and C. Melhuish, 85–96. Springer, 2014.

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 it must 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 the quality of human-machine interactions, human approval and appreciation of the A/IS, trust in the A/IS, adaptability of the A/IS to human users, and human benefits in the presence or under the influence of the A/IS. A range of ethical/normative aspects to be considered can be found in the UK standard on Robot Ethics (BSI, 2016). 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, 2001), 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 autonomous systems, including robots (Fisher, Dennis, and Webster, 2013).

Evaluation may be done by first parties (designers/manufacturers, and users) as well as third parties (e.g., regulators or independent testing agencies). 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.

Candidate Recommendation

Evaluation must be anticipated during a system's design, incorporated into the implementation process, and continue throughout the system's deployment. 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.

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

Background and Analysis

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 may not have *identical* sets of norms. People will have some unique expectations for humans than they do for machines (e.g., norms governing the regulation of negative emotions, assuming that machines do not have such emotions), and people will have some unique expectations of A/IS that they do not have for humans (e.g., that the machine will sacrifice itself, if it can, to prevent harm to a human).

Candidate 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 and Analysis

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 (e.g., a passport application AI rejected an Asian man's photo because it insisted his eyes were closed; Griffiths, 2016); information processing (e.g., speech recognition systems are notoriously less accurate for female speakers than for male speakers; Tatman, 2016); decisions (e.g., a criminal risk assessment device overpredicts recidivism by African Americans; Angwin, et al., 2016); and even in its own appearance and presentation (e.g., the vast majority of humanoid robots have white "skin" color and use female voices) (Riek and Howard, 2014).

The norm identification process detailed in Section 1 is intended to minimize individual designers' biases, because the community norms are assessed empirically. The process also seeks to incorporate values and 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 AI systems and integrating community outreach into the evaluation process (e.g., DO-IT program; RRI framework). Behavioral scientists and members of the target populations will be particularly valuable when devising criterion tasks for system evaluation. Such tasks would assess, for example, whether the A/IS applies norms in discriminatory ways to different races, ethnicities, genders, ages, body shapes, or to people who use wheelchairs or prosthetics, and so on.

Candidate Recommendation

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

Issue 3:
Challenges to evaluation by third parties.

Background and Analysis

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 an A/IS's 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). 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 it matches its normative requirements (e.g., acting in accordance with relevant norms; recognizing other agents' norm violations).

These different evaluation techniques can be assigned different levels of "strength" — strong ones demonstrate the exhaustive set of an A/IS's allowable behaviors for a range of criterion scenarios; weaker ones sample from criterion scenarios and illustrate the system's behavior for that subsample. In the latter case, confidence in the A/IS's 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 itself) 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 and who has the right to correct the systems, or also which kind of A/IS should be subject to accountability requirements.

Candidate Recommendation

To maximize effective evaluation by third parties (e.g., regulators, 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 possibly achieve accountability to the relevant communities.

References

- Angwin, J., 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.
- British Standards Institution. BS8611:2016, "Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems," 2016.
- Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues. FTC Report." Washington DC: Federal Trade Commission, 2016.
- Fisher, M., L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56 (2013): 84–93.
- Griffiths, J. "New Zealand Passport Robot Thinks This Asian Man's Eyes Are Closed." *CNN.com*, December 9, 2016.
- Riek, L. D., and D. Howard. "A Code of Ethics for the Human-Robot Interaction Profession." *Proceedings of We Robot*, April 4, 2014.
- Tatman, R. "Google's Speech Recognition Has a Gender Bias." *Making Noise and Hearing Things*, July 12, 2016.

Further Resources

- Anderson, M., and S. L. Anderson eds. Machine Ethics. New York: Cambridge University Press, 2011.
- Abney, K., G. A. Bekey, and P. Lin. Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, MA: The MIT Press, 2011.
- Boden, M., J. Bryson et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science* 29, no. 2 (2017): 124–129.
- Coeckelbergh, M. "Can We Trust Robots?" *Ethics and Information Technology* 14 (2012): 53–60.

- Dennis, L. A., M. Fisher, N. Lincoln, A. Lisitsa, and S. M. Veres. "Practical Verification of Decision-Making in Agent-Based Autonomous Systems." *Automated Software Engineering* 23, no. 3, (2016): 305–359.
- Fisher, M., 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 (2016): 114–137.
- Fleischmann, K. R. *Information and Human Values*. San Rafael, CA: Morgan and Claypool, 2014.
- Governatori, G., and A. Rotolo. "How Do Agents Comply with Norms?" in *Normative Multi-Agent Systems*, edited by G. Boella, P. Noriega, G. Pigozzi, and H. Verhagen, Dagstuhl Seminar Proceedings. Dagstuhl, Germany: Schloss Dagstuhl — Leibniz-Zentrum fuer Informatik, 2009.
- Leet, E. H., and W. A. Wallace. "Society's Role and the *Ethics of Modeling*," in Ethics in Modeling, edited by W. A. Wallace, 242–245. Tarrytown, NY: Elsevier, 1994.
- Jarvenpaa, S. L., N. Tractinsky, and L. Saarinen. "Consumer Trust in an Internet Store: A Cross-Cultural Validation." *Journal of Computer-Mediated Communication* 5, no. 2 (1999): 1–37.
- Mahmoud, M. A., M. S. Ahmad, M. Z. Mohd Yusoff, and A. Mustapha. "A Review of Norms and Normative Multiagent Systems." *The Scientific World Journal*, (2014): 1–23.