From Machine Ethics to Internet Ethics: Broadening the Horizon

Pradeep K. Murukannaiah
Delft University of Technology

Munindar P. Singh
North Carolina State University

***Abstract*—This article introduces some of the key concepts and challenges pertaining to ethics from the standpoint of Internet applications.**

■ **Ethics is an** ancient concern going back to the earliest works in literature and philosophy in just about any culture. Ethics, at its core, is all about understanding the effects of one individual's actions toward another.[7] That is, one's ethical actions help others and unethical actions harm others (see the sidebar on ethics for a longer discussion).

Machine ethics, a line of inquiry in ethics that has gained attention,[2] deals with the ethics of machines. Its goal is to characterize or build machines that can act ethically. As an academic discipline, machine ethics is concerned with both metaethical and practical normative questions. An example of the former is can ethics ever be computed at all? As computer scientists, we would generally assume yes. An example of the latter is what principles and procedures are required to program machines to act ethically? Machine ethics is typically distinguished from *computer ethics*, which is concerned with the ethics of humans using computers and especially of computing professionals in building and deploying software.

Moor[5] argues that we should study machine ethics because 1) we want machines to treat us well (as machines become more sophisticated and autonomous, the stronger this motivation becomes), and 2) we can better understand (human) ethics by trying to systematically instill ethics in machines. However, why should we separate human and machine ethics?

Digital Object Identifier 10.1109/MIC.2020.2989935
Date of current version 21 July 2020.

SIDEBAR: INTERNET ETHICS: FAIRNESS, ACCOUNTABILITY, TRANSPARENCY

■ **Internet applications are** increasingly based on artificial intelligence. Some applications involve explicit and visible intelligent, networked agents—ranging from personal assistants on smart phones to self-driving cars on streets. Other applications do not involve direct action but provide decision assistance, such as in providing judges sentencing guidelines, e.g., to set prison terms for someone convicted of a crime. Still other applications involve uses of intelligence hidden in the services, such as Facebook and Amazon, that determine what news we see and what products we buy. The extraordinary characteristics of increasingly intelligent Internet applications, including the detailed access they have to our personal data and the fine-grained control they can exert on our lives, have been a cause of growing concern. Consequently, Internet ethics is garnering increasing attention, and rightly so. This new department will focus on some of the key societal aspects of Internet Computing, specifically, the challenges that come to the fore as artificial intelligence technologies become dominant in virtually all applications of computing. It will take a broad sociotechnical approach to ethics, incorporating concerns that arise from deployed Internet applications such as fairness, accountability, and transparency. The upcoming articles in this department will delve into these topics in greater detail.

Machines, of course, are an indispensable part of almost every human activity. And they are increasingly powerful. However, it is important to distinguish *autonomy* from *automation*. Machines demonstrate automation in that they can carry out complex feats of reasoning. But we characterize autonomy in human and societal terms. A machine acts on behalf of a human and therefore reflects the autonomy of that human. For example, if your agent transfers your money to a Nigerian Prince, it is your money that is lost, not the agent's—what would that even mean? Your bank allows the transfer only because the transfer was performed on your behalf. If a smart speaker records your private conversations, again we do not accuse the box of malfeasance but the people and organizations on whose behalf it acts.

In other words, although machines are increasingly powerful, they act not in artificial societies consisting only of machines, but in hybrid societies of humans and machines. Then, ethics being a study of how one's actions affect another, we posit that the ethics of humans and machines are deeply intertwined.

This (human and machine) integrated view of ethics is what we envision as *Internet ethics*. Specifically, we bring forth ethics in the domain of Internet applications. Elements of the domain that are relevant include people (users, developers, and administrators), machines (computers as well as smart devices on the IoT), and resources (data and services).

Besides the traditional resource conflicts in many ethically charged situations, in Internet apps, unethical outcomes may arise due to complexity (of the tools), malice (security attacks), lack of confidence by users (due to lack of explicability and transparency), and biases in data and reasoning.

WHERE DOES INTERNET ETHICS APPLY?

Let us consider a sampling of Internet application scenarios where ethics is crucial.

Traffic

A city obtains data from sensors about the number of vehicles at different traffic signals in the city at different times in a day. What measures can the city take to reduce congestion on its roads?

At one level, this is a classical optimization problem—adjust traffic signal durations at selected junctions or broaden certain roads to maximize traffic flow, minimizing congestion. What makes this an ethical problem lies in understanding why congestion is a problem to the residents of the city and how potential interventions affect their lives. Suppose the residents value the environment. Then, a solution to the congestion problem may be not only about optimizing traffic flow but also about enriching public transportation and planning city services to reduce load on the transportation system. Alternatively, suppose the residents value spending time with their families. In this case, the solution should emphasize city planning, increasing residential places near work places so that people can minimize commute to work. In essence, the focus should not simply be on the cars on roads causing congestion but on how people are affected by congestion. Such analyses, which must be *transparent* and require an understanding of the *values* of people, are an integral part of ethical reasoning.

Ambulances

Multiple ambulances, carrying patients, are currently on the road distributed across a city. How can the ambulances collectively decide on which ambulance goes to which hospital?

The fact that the ambulance allocation affects *human lives* makes it a typical ethical problem. A number of ethical concerns, including *transparency* and *explicability* are crucial in this scenario. For example, the system should be able to explain why an ambulance went to a hospital A that is far from the ambulance's current location than to another hospital B that was nearer—may be the route to hospital A was more congested; may be hospital A did not have adequate facilities to treat the patient in the ambulance; or, maybe two ambulances were at about the same distance from hospital A and only the one carrying the more severe of the two patients can be sent to hospital A. Meaningful human *control* is an important ethical consideration in this scenario. The medical professionals on board the ambulance should be able to override the recommended allocation and choose to take the patient to an alternative hospital.

Policing

A city's police department receives reports of multiple incidents at the same time. Given resource constraints, in what order should the department investigate the incidents?

This scenario is obviously loaded with ethical concerns. The police department's resource allocation algorithms should *not be biased* in that they must treat various communities fairly. Modern societies generally offer legal protection against discrimination on the basis of race and gender. However, as is well-known, innocuous attributes (e.g., post code) can serve as surrogates for sensitive attributes (e.g., race and wealth). Societal *control* is thus important in this scenario—police officers may exercise some discretion in what incident to investigate more urgently but would remain *accountable* for their decisions.

Power Usage

The residents of a neighborhood can time-shift some of their electrical load by running appliances such as clothes washers and dryers at specific times. How can these residents collectively shift their loads to reduce the peak aggregate demand on the electrical grid while fulfilling everyone's needs? (Reducing peak demand has benefits to sustainability by reducing the need for power plants.)

Ethical considerations in this scenario include the need for residents to be *prosocial* (being considerate to each other by being flexible) and being *truthful* in reporting their preferences to each other.[11] In particular, those who are willing to be flexible for the greater good should not have to carry an unfair share of the burden.

Smartphones

A smartphone user, Sam, finds it cumbersome to update his phone's ringer settings, such as its loudness, multiple times a day. Without appropriate settings, Sam misses important calls because his phone is silent when it should not have been or finds himself in awkward situations because his phone rings loudly when it should be silent. How can a smart ringer app automatically adjust the phone's ringer settings?

Although the smart ringer app's actions seem innocuous, the app must reason about a number of ethical concerns. Consider, for example, that Sam is in a library and is receiving a call. Now, is it ethical for the phone to ring loudly on a casual call from a friend? Alternatively, is it ethical for the phone to stay silent on an important call, e.g., from Sam's spouse, who needs immediate assistance? If the phone, indeed, rings loudly, is it ethical for others in the library to judge Sam as inconsiderate of others? Perhaps, the app can send an explanation to others in the library as to why it had to ring loudly.[1] In that case, is it ethical for the app to compromise Sam's privacy so as to not harm his social reputation?

SIDEBAR: ETHICS

Ethics is the study of an individual's *conduct* or way of acting in a society, where the actions taken by the individual affect the outcomes enjoyed by others. Three main subject areas in the field of ethics are *normative ethics*, *metaethics*, and *applied ethics*.[4]

Normative ethics is the study of practical means ("practical" here indicating action by an individual or society) of determining the right or wrong of one's conduct. Thus, normative ethics is concerned with principles and guidelines of what one ought to do, ethically, under different circumstances.

There are three main classical theories in normative ethics: *virtue*, *deontological*, and *consequentialist*. The virtue and deontological theories focus on the action, itself, to determine its ethicality, i.e., its rightness or wrongness. In virtue theory, the ethicality of an action comes from the inherent character (virtues) of an individual. In contrast, in deontological theory, ethical actions are those that conform to rules, laws, and norms. In contrast to virtue and deontological theories, in consequentialist theory, the consequences of an action determine its ethicality. Specifically, *utilitarianism* is a consequentialist theory in which ethical actions are those that maximize utility for everyone affected by the action.

Consider an example to understand the differences in the reasoning processes the three normative ethical theories advocate. Should an ambulance with one patient but one vacant spot stop to help at an accident when en route to a hospital? It is virtuous to help those in need. From a deontological perspective, the rule can be that a healthcare professional cannot let a patient die on the road without trying to help. Finally, from a consequentialist perspective, suppose the risk is high if the patient on board cannot get to a hospital within 20 min. So the ethical course would be to save the patient on board and let some other ambulance take care of the patient lying on the road.

Metaethics is the study of the very nature (i.e., meaning, origin, and basis) of ethical concepts, judgements, and propositions. It incorporates *metaphysical* questions, such as can ethics exist independently of humans? And, it incorporates *psychological* questions such as what is the mental basis for an ethical judgement?

Applied ethics is the study of ethics within a specific domain or a context. Examples of applied ethical fields include business ethics, medical ethics, military ethics, and so on.

Machine ethics,[2] as a branch of applied ethics, is concerned with developing machines such as robots that act ethically. Moor[5] recognizes four varieties in which machine ethics can manifest itself: 1) *ethical-impact agents*, whose actions have ethical implications (similar to the consequentialist ethics), whether intended so or not; 2) *implicitly ethical agents*, whose actions are ethical although they are not explicitly programmed to be ethical (however, they may have been programmed to avoid unethical actions); 3) *explicitly ethical agents*, which explicitly represent and reason about ethics in choosing their actions; and 4) *fully ethical agents*, akin to humans, have metaphysical characteristics (such as intentionality and consciousness) and are capable of explicit ethical judgements and are competent in justifying those judgements. Moor recognizes that the (metaethical) question as to whether or not a machine can be a fully ethical agent may not be resolved in the near future but identifies realizing explicit ethical agents as a challenging and worthwhile pursuit.

UNDERSTANDING ETHICS IN STS

Our conception of Internet ethics builds on the idea of *sociotechnical systems* (STSs).[9] An STS includes humans and organizations as social entities and agents (which abstract over computing

SIDEBAR: VALUE THEORY

Whereas philosophers have traditionally approached ethics from the standpoint of actions and decision making, which are often described in toy examples bereft of context, works in the social sciences consider how ethical considerations play out from a more personal and social perspective.

A central construct in this setting is of *values*. Social psychologists understand values as 1) deeply held beliefs and preferences of people that motivate actions and 2) universally valid constructs. They have proposed competing lists of values that are understood as cutting across application domains and cultures. For instance, a leading theory of values from Schwartz[8] recognizes ten universal values: *self-direction*, *stimulation*, *hedonism*, *achievement*, *power*, *security*, *conformity*, *tradition*, *benevolence*, and *universalism*.

Schwartz argues that the values listed above are universal because each of those is derived from one or more universal requirements of human existence, including the needs of: 1) individuals as biological organisms, 2) coordinated social interaction, and 3) survival and welfare of groups. The values may mutually conflict or not. For example, self-direction and conformity conflict, but conformity and tradition are compatible. Schwartz models the relationships among the ten universal values along two dimensions. The ends of the first dimension are self-enhancement (including achievement) and self-transcendence (including benevolence). And, the ends of the second dimension are openness to change (including self-direction) and conservation (including tradition).

It is often appropriate to consider an expanded set of value-based constructs to capture the needs of an application. For example, we may consider *privacy* as derived from self-direction and *safety* from security. Users may have personal preferences between value-based constructs—for example, Alice may prefer safety to privacy, and Bob may prefer privacy to safety.

Understanding the values at stake in a decision context is an important first step in employing ethical reasoning in that context. Suppose a decision is to be made regarding whether to install a camera in a park that Alice and Bob frequent. Alice may find cameras that capture what people do in a park legitimate (and hence ethical) whereas Bob may find the resulting data gathering unethical.

entities) as technical entities. Each agent acts on behalf of a human,[6] and together, they form what we call a *human–agent duo*. The duo conception highlights that the effects that the participants in an STS perceive are aggregated from the joint work of each participating human and that human's agent.

The Internet provides an infrastructure on which humans and agents form duos and how the duos participating in an STS interact. The Internet also provides resources such as data and services to facilitate ethical reasoning by the human–agent duos.

A human–agent duo can participate in multiple STSs. Indeed, each Internet application can be modeled as an STS. For example, there can be an STS of ringer app users, an STS of the residents of a neighborhood, or an STS of emergency medicine (which includes hospitals, ambulances, doctors, patients, ambulance drivers, and so on).

Figure 1 shows schematically how key representations and processes can be allocated to STSs and to human–agent duos, and how these line up with each other. The participants in an STS influence the STS while the STS informs the participants' decision making. Importantly, an STS is not a separate computational entity but is realized through the human–agent duos that constitute it, and which it encompasses. An STS may be engineered or emergent, and for Internet applications would usually have both engineered and emergent aspects.

A human–agent duo's reasoning process captures its concrete decision making. As stated above, this process is informed by the STSs in

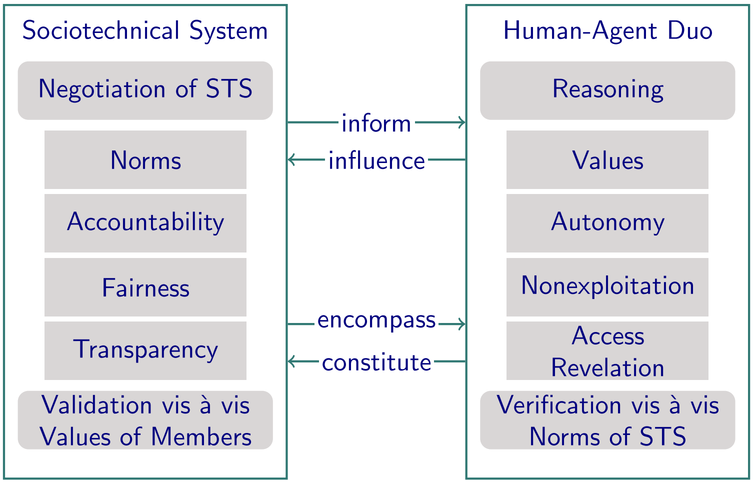

Figure 1. Relating human–agent duos and sociotechnical systems (STSs). The processes (boxes with rounded corners) and representations (boxes with sharp corners) shown at matching levels portray a duality between the STS and its constituent human–agent duos with respect to ethical reasoning.

which the duo participates. A duo's verification process checks decisions by it (and by other duos) with respect to the norms of the relevant STSs.

The negotiation process of an STS captures how it is constructed by its participants. Likewise, the validation process concerns how an STS is aligned with the values of its participants. Usually, validation and negotiation would be interleaved processes that characterize the lifecycle of an STS.

Figure 1 shows the duality between STSs and human-agent duos that ethical reasoning brings forth. The values (including value preferences) of a human govern the associated duo. The norms of the STS operationalize the collective values of its constituent duos. A duo is autonomous in an STS exactly to the extent that it is accountable for its actions. The capability of a duo to neither exploit nor be exploited corresponds to the STS demonstrating the property of fairness. Likewise, a duo's capability to reveal its representations and reasoning and access the representations and reasoning of other duos corresponds to the STS demonstrating the property of transparency.

At a deeper level, values and norms reinforce each other—norms emerge from values and the established norms influence an individual in developing value preferences. This duality can also be observed in Schwartz's value model (see the sidebar on value theory). Self-direction, which motivates an individual to develop his or her own value preferences, is a universal value; so is conformity, which motivates an individual to conform to established norms.

A sociotechnical conception of ethics is a significant departure from the single-machine view of ethics prevalent in the computing literature. As a case in point, *algorithmic fairness* concerns making a single algorithm fair.[10] For example, is a predictor fair in the predictions it outputs? However, this narrow, statistical view of fairness can potentially subvert actual fairness in real life.[3] To counter such limitations, algorithmic fairness must be understood and tackled from a holistic, sociotechnical, perspective that tackles questions such as the following. For example, what biases do the processes that collect the data on which an algorithm (whose fairness is under scrutiny) is trained, have? What biases do decision makers such as judges deciding on a convict's parole have in configuring these algorithms? Do they choose appropriate decision thresholds? Do they correctly interpret the recommendations the algorithms provide? And, do decision makers ascribe greater confidence to an algorithm's recommendations than is merited?

CONCLUSION

The emphasis on Internet ethics makes ethical inquiry, already a fascinating subject, even more fascinating. In essence, Internet ethics is not merely a branch of applied ethics, but it broadens the horizon of ethics in all directions.

From the standpoint of practice, the examples we discussed demonstrate that Internet ethics applies in a variety of situations, including smartphone apps that automate day-to-day activities, smart city applications that enhances the quality of life, and resource allocation problems (e.g., scheduling ambulances) where human lives are at stake.

From the standpoint of theory, Internet ethics demonstrates that ethics applies not merely to the decision making of individuals but more importantly also to the sociotechnical settings in which they function: in terms both of the social and organizational structures where their decisions have consequence and of the technical entities that facilitate or enable different decisions.

Subsequent articles in this department will explore this topic from diverse perspectives matching the diversity of Internet applications.

REFERENCES

1. N. Ajmeri, H. Guo, P. K. Murukannaiah, and M. P. Singh, "Designing ethical personal agents," *IEEE Internet Comput.*, vol. 22, no. 2, pp. 16–22, Mar. 2018.
2. M. Anderson and S. L. Anderson, editors, *Machine Ethics*. Cambridge, UK: Cambridge Univ. Press, 2011.
3. S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," 2018. [Online]. Available: https://arxiv.org/abs/1808.00023
4. J. Fieser, "Ethics," in *The Internet Encyclopedia of Philosophy*, J. Fieser and B. Dowden, Eds., 2020. [Online]. Available: https://www.iep.utm.edu/ethics/
5. J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 18–21, Jul. 2006.
6. P. K. Murukannaiah, N. Ajmeri, C. M. Jonker, and M. P. Singh, "New foundations of ethical multiagent systems," in *Proc. 19th Int. Conf. Auton. Agents MultiAgent Syst.*, May 2020, pp. 1–5.
7. R. Paul and L. Elder, *The Thinker's Guide to Ethical Reasoning: Based on Critical Thinking Concepts & Tools*. Lanham, MD, USA: Rowman & Littlefield, 2019.
8. S. H. Schwartz, "An overview of the Schwartz theory of basic values," *Online Readings Psychol. Culture*, vol. 2, no. 1, pp. 11:1–11:20, 2012.
9. M. P. Singh, "Norms as a basis for governing sociotechnical systems," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, pp. 21:1–21:23, Dec. 2013.
10. S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. Int. Workshop Softw. Fairness*, 2018, pp. 1–7.
11. G. Yuan, C.-W. Hang, M. N. Huhns, and M. P. Singh, "A mechanism for cooperative demand-side management," in *Proc. 37th IEEE Int. Conf. Distrib. Comput. Syst.*, Jun. 2017, pp. 361–371.

Pradeep K. Murukannaiah is currently an Assistant Professor with the Interactive Intelligence Group, Delft University of Technology, Delft, Netherlands. He received a Ph.D. degree in computer science from NC State University, Raleigh, NC, USA. The overarching theme of his research is engineering socially intelligent applications. Contact him at p.k.murukannaiah@tudelft.nl.

Munindar P. Singh is currently a Professor in Computer Science and a Co-Director of the Science of Security Lablet, NC State University, Raleigh, NC, USA. He received a Ph.D. degree in computer sciences from The University of Texas at Austin, Austin, TX, USA. His research interests include the engineering and governance of sociotechnical systems, and in AI ethics. He is an IEEE Fellow, a AAAI Fellow, and a former Editor-in-Chief of *IEEE Internet Computing* and *ACM Transactions on Internet Technology*. He is the corresponding author of this article. Contact him at m.singh@ieee.org.