You, Me, or Us: Balancing Individuals' and Societies' Moral Needs and Desires in Autonomous Systems

Joseph A. Blass (Northwestern University; joeblass@u.northwestern.edu)
DOI: 10.1145/3175502.3175512

Abstract

Autonomous systems are increasingly taking actions with moral consequences. While certain universal norms may be built into such systems, other areas should be personalizable to the end user. This essay motivates this problem, explores the ways such systems are already making decisions with moral ramifications, and proposes a path forward.

Introduction

We are on the cusp of a new era, in which autonomous systems make decisions with moral consequences that impinge on numerous aspects of our lives. AI programs ought to consider the moral ramifications of their actions (Scheutz, Malle, & Briggs, 2015), but can designers engineer for such systems a single universally standard set of morals to cover all situations? It's unlikely: while the law takes a stand on certain matters, others are left to individuals, whose ethical and moral norms vary within and across societies. Any complete set of morals accepted by one person will somehow be in violation of another's. We equally cannot avoid building any ethical or moral reasoning capabilities into autonomous systems, or let them do whatever humans tell them: such systems could do great harm, either through ignorance or by human instruction. The core ethical challenge facing AI researchers and engineers is balancing individual ethical and moral preferences with the norms needed for society to function.

This essay will argue why there are many situations in which individuals should have some control over their machines' ethics. It will argue that the solution to this problem is to collectively agree upon a set of universal principles for autonomous systems to abide by, but which will cover *only a portion* of all possible ethical and moral situations, with users having some say over the rest. These built-in norms will be those which are core to society and necessary for humans to trust these systems. Building such a system will be time-consuming. It will involve philosophers, ethicists, sociologists and psychologists working to determine our minimal collective norms; AI experts to implement them; and lawyers and lawmakers to write the laws regulating them.

Morals Vary Within & Across Societies

Morals are principles cleaving right from wrong, and ethics are codes of conduct based on morals. While these are fundamentally different, this essay treats these terms as interchangeable as they relate to AI. Human morality is partly rooted in emotion and instinct (Haidt, 2001); fully modeling it may prove to be an AI-complete problem. Constructing ethical norms that lead to decisions humans consider moral, however, is within the purview of current AI technologies.

While humans may generally agree on some moral norms (e.g., don't steal from or attack other members of society), there is no one universal moral system that we all subscribe to. The specific morals and ethics we develop are defined in large part by culture and experience. One well-studied example distinguishes between cultures that value individualism vs. collectivism. Broadly speaking, Individualist (generally Western) societies focus on individual actions and value personal freedoms, whereas Collectivist (generally Eastern) societies focus on social outcomes and value social harmony. Societies on different sides of this fault line show different patterns of moral judgment. In one case, researchers asked Canadian and Chinese children to evaluate lies that either helped or hurt a social group (Lee, Cameron, Xu, Fu, & Board, 1997). Chinese children judged prosocial lies more positively than antisocial ones, but Canadian children did not. For the Chinese children the relevant feature was the effect of the action on the group, whereas for the Canadian chil-

dren it was the nature of the action itself. Cultural moral differences like these are not easily dismissed, and it would be wrong for engineers from one culture to force members of another to use machines that abide by the culturally-specific ethical conventions of the first. (For a review of the role of culture in moral judgments, see (Sachdeva, Singh, & Medin, 2011)).

Even *within* societies, people vary dramatically in their moral judgments. Psychologists can identify overall trends in patterns of responses, but rarely find universal agreement. For example, Haidt and colleagues (Haidt, Koller, & Dias, 1993) asked participants in the US and Brazil whether violations of moral norms around purity (e.g., using a national flag to clean a toilet) were permissible. Most people in both countries identified these actions as moral violations, but rated them as being harmless. However, they disagreed about whether these actions should be permitted, and socio-economic status predicted their responses more than nationality. If people disagree within a society about whether certain types of harmless actions are permissible, it is hard to imagine a standards board determining a complete set of principles that all users will perceive as morally just without being overly restrictive or permissive.

Some may say harmless actions should not be regulated, but an AI system acting on a person's behalf should not take actions that are offensive, even if they cause no concrete harm. Robots shouldn't use flags as rags. Moreover, people disagree over how wrong certain unambiguously harmful actions are. In the classic trolley problem, a trolley will hit five people, who can be saved by sacrificing one person. In the *switch* scenario, the trolley can be turned onto a side track with one person (who will die). In the *footbridge* scenario, a person can be pushed in front of the trolley, stopping it. People overwhelmingly say taking action in the switch scenario is better than taking action in the footbridge scenario, but some think acting in the footbridge scenario is permissible. Fully 60% of participants in a recent study endorsed acting in the footbridge scenario (compared to 79% for the switch scenario) (Hristova & Grinberg, 2016). Response patterns change if a humanoid robot is acting (the switch scenario is more permissible) and change again when a non-humanoid automated system takes action (both scenarios are more permissible). Some people see actions as more or less permissible depending on the identity of the victims (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009). Clearly the reliable trends these studies reveal do not point to a universal "right" or "wrong" answer to these scenarios.

Certainly a society can forbid autonomous machines from taking certain types of action. Prohibiting unprovoked aggression and violence, theft, and abuse are obvious restrictions to place on computers. We have standards of psychopathy among other mental illnesses, and we should not build AI systems that meet their diagnostic criteria. Furthermore, human history and societies are filled with instances where powerful subpopulations claimed a moral right, if not imperative, to oppress and exploit other subpopulations. AI systems should never be allowed to promote oppression.

AI should also not be used to enable human hypocrisy. For example, Bonnefon and colleagues found that most people say autonomous vehicles should minimize loss of life on roads, even if doing so involves sacrificing the driver. These same people indicated that they would not want to buy such a self-driving car (Bonnefon, Shariff, & Rahwan, 2016). They are happy to impose on others an ethical system that they themselves refuse to follow. Nevertheless, we recognize that this desire cannot be satisfied in our interactions with other humans: we don't expect our doctor to sacrifice her other patients to save us. When deciding what norms to build into an autonomous intelligent system, people similarly must recognize that those ethical norms may limit the actions a machine can take on their behalf.

Nevertheless, there is a wide range of circumstances in which people should have some say over the behavior of their machines. These circumstances arise outside of those situations involving the core moral norms that protect society, and concern both positive and negative situations: not only those where someone is likely to get hurt, but also ones where the system will have to decide who, and how, to help. While this may sound like the

stuff of science-fiction, AI systems are already making these kinds of decisions, and it is problematic that they are not doing so with an explicit and consistent set of moral principles.

Moral Hazards for Autonomous Systems

Much ink has been spilled over the ethics of self-driving cars, but many of the trickiest and most difficult ethical questions arise with other, more mundane systems that are already in widespread use and affect our lives. One such class of system is machine-learning based law enforcement assistants. These are being used to predict flight and recidivism risk when determining bail at trial, and for facial recognition purposes for law enforcement. Since machine learning algorithms learn from their inputs, if the input encodes systemic bias, the algorithms will too. One recidivism predictor gives harsher scores to African-American defendants than to White ones, despite race not being an input criterion (Angwin, Larson, Mattu, & Kirchner, 2016). Criminal facial-recognition software disproportionately mistakenly identifies African-Americans, who are more likely to have had contact with the police, regardless of criminality (Garvie, 2016). Being mistakenly called in for questioning or unfairly denied bail can have a devastating effect on a person's life.

Reducing the bias within those systems is difficult (Diakopoulos, 2016), but if bias cannot be eliminated, the systems should not be used. If bias is eliminated, societies still must decide how confident their system's judgments must be. A system that requires a lower confidence threshold will miss fewer criminals but flag more innocents; a system with a higher threshold will do the reverse. Should the system be more concerned with security or liberty? This is a choice that must be made *by* the community, not *for* the community by the software company.

Or consider automated hospital management systems. These systems, which are in increasingly widespread use, perform a variety of operations including bed assignment and supply management. Both of these tasks have potential moral implications. In the midst of an epidemic, a hospital may run out of beds. On what basis will an automated system determine who gets a bed and who does not? How quickly (and which) patients that have beds will be discharged to make room for others? These decisions vary by hospital, by patient, and by epidemic. Doctors of course have the final say, but a well-designed automated system ought to facilitate that decision-making process. For doctors and administrators to rely on the system during a hectic situation, they must trust it to make the right decisions. This trust requires transparency and the knowledge that the system will implement the hospital's priorities.

Stocking medicines costs money. The factors that determine how much medicine of any kind to stock are particular to individual hospitals. What happens in the rare instances when they have more patients than medicine? Though unlikely, these situations are nearly guaranteed to occasionally occur, for example in the early stages of an epidemic. In such cases, which patient gets the medicine, and who must wait or get another treatment? Even using a first-come-first-serve basis is making an ethical choice. A patient might make a different decision about which hospital to go to if they think they are more likely to get a rare medicine at a hospital with a different disbursal process. Again, hospital administrators must understand and control how their systems make these decisions.

Automated hospital systems solve more problems than they create. They are more efficient and accurate, integrate more data, and are less biased than their human counterparts (i.e. they don't care whether a patient is rude). The point is not that they will be immoral or unethical, but that hospitals deal with ethical grey areas and develop their own ethical standards within clearly defined professional ethical standards. Decisions made by automated systems should be consistent with hospitals' developed norms.

This example also shows that a decision about who to hurt can be a decision about who to help. It is a truism that we cannot help everybody. A self-driving car may do the most good by abandoning its owner and driving to the country with the highest rate of malnourishment to volunteer itself for Meals on Wheels, but few would argue we should build such cars. Resources are limited, and AI systems

need to know who to help, and how. They must express positive values, not only avoid negative ones.

There is at least one product on the market which people will soon want to have express positive moral and ethical values, including ones they should have control over: Mattel's Hello Barbie. Hello Barbie is a doll with AI-driven conversational abilities that learns about your child. Hello Barbie received negative attention over security issues (it requires an internet connection and does all its computation on the cloud), but a Hello Barbie of the near future may run locally and avoid these concerns. Nonetheless even without any security concerns, such a system potentially involves a range of moral concerns, both positive and negative.

The largest of these is the moral development of the child. Children are constantly learning, and they learn how to be members of society through social interaction. We don't know how interacting with such a doll will affect a child's social and moral development, but children may well learn from their interactions with it. Whatever values the doll has been instilled with may potentially be learned by the child. Again, the doll should discourage harm, theft, etc. But what about the benevolent lying example discussed above? This (among others) would presumably be a value the child's parents would want to teach them. If the parents are teaching the child that otherwise harmless prosocial lies that promote social cohesion are OK, but the doll teaches the child that lying is never OK (even to protect someone's feelings), then the toymaker is directly undermining the parents' moral instruction. And what about religion? A parent raising a child in one religion might object to the doll expressing beliefs in another religion (or none), and a parent in an atheist household might object to the doll expressing any religious beliefs whatsoever. Giving the parents some control over what the doll teaches the child, control they have in other domains such as what media the child has access to, will be crucial. (On the other hand, society will retain an interest in other areas, such as preventing parents from making the doll teach their child bigotry.)

What should the doll report or keep secret? If the doll detects signs of depression, should it tell the parents? What if a child with homophobic parents tells the doll she is gay? What if the doll detects signs of abuse? These are all things modern AIs could be trained to glean from conversation. Reporting depression might help the child, but reporting that the child is gay could hurt her, depending on how such news is received. If the doll reports on abuse to the government, then the doll is actively surveilling its owners; if it does not, then the doll-makers have created a system that can detect abuse, but ignores it. Furthermore, different cultures have different standards of abuse: in many modern cultures striking a child is always considered abusive; in others, corporal punishment is a widely used and socially acceptable form of discipline. Corporal punishment may well fall into the category that we collectively decide is never acceptable. If so, people in societies that routinely use corporal punishment are more likely to be reported if they own the doll: they will be punished for having purchased it. And it is easy to take this benevolent principle to an absurd Orwellian extreme, with the doll being made to encourage a child to report her dissident parents. The tension between encoding the ethics necessary to maintain society and enabling individuals and organizations to teach their AIs their own ethical systems is as much about restraining society to protect individuals as it is about restraining individuals to protect society.

In some domains, current technologies already rely on default strategies that may have moral consequences. Non-prosocial lying is such a case: administrative assistants, for example, may lie about their bosses' availability; should a digital assistant be able to tell such lies? If so, to what degree can the lies be taken? Though not in widespread use, digital secretaries are already being used in the real world (e.g. (Pejsa et al., 2014)), and need an answer to this question. To take another example, should Siri try to stop you from drunk-dialing your ex (or at least argue with you about it)? When should a digital assistant make decisions for you, or try to impact your decision-making?

Towards Ethical AIs

It should now be clear that there is no single morality, either across or within cultures, that could be built into an automated system, and that humans should have some say in their machines' ethics. However, we cannot simply trust humans to make their machines always do the right thing. People convicted of driving drunk have breathalyzers installed in their cars that lock the engine; people with restraining orders could have software installed on their phones to prevent harassment. While in general you have the right to regulate your own behavior, and your technology should help you do that, in some cases, society (or the state) clearly has an overriding interest that limits your behavior. Humans are all too happy to exploit and harm each other, especially when cloaked in anonymity or acting on someone who is not a member of their in-group. It is not only psychopaths or evil people who behave this way; it is well known that the internet can bring out the worst in people (Suler, 2004), and AI should not help them. Again, the pressing issue is neither determining the set of universal morals nor building a fully personalizable ethical system. Rather, the challenge involves navigating the tension between imposing an external ethical system upon people and protecting the interests of society as a whole. The rest of this essay will deal with how to address this problem.

The first step is to collectively determine the set of ethics which *must* be built in to protect us from psychopathic AIs and the worst of human behavior. These ethics will by definition not cover all situations, but only those which society has a fundamental interest in regulating. Existing laws can be the starting point for this discussion: if we have collectively decided humans should not be allowed to do something, computers should not either. Of course, laws trade off against each other, and there are situations wherein otherwise forbidden things are permitted (e.g. violence in self-defense). This process will bring together scholars of the humanities, social sciences, and law to determine this core set of ethical principles, and AI researchers to explain how the realities of implementation will affect the theory of what is being implemented (and to implement it). Everyone involved must understand and communicate that the ethics being agreed upon will not only be regulating systems used by others, but by ourselves, given the problem that (Bonnefon et al., 2016) identify. The ethics defined for autonomous systems may well be different from those governing humans (Hristova & Grinberg, 2016); the point is that we must be clear about what it is we are building, and why.

An approach like Conditional-Preference Nets (CP-Nets, (Greene, Rossi, Tasioulas, Venable, & Williams, 2016)) may work for this task. CP-Nets define what behaviors are preferred under particular conditions. For example, in a situation where a person is in immediate danger, having a robot help that person escape might be preferable to violently defending them, but attacking the aggressor might be permitted if escape is impossible (unless, for example, the aggressor is a law enforcement officer). If the person is being intimidated without direct threat, however, the robot is not allowed to attack the aggressor until the threat becomes tangible. CP-Nets can encode these contextual differences in preferences.

Once that work is done, we must determine a framework within which other ethical concerns can be identified and personalized to the end-user, including the domains, the means, and the extent to which they can be personalized. Such personalization will take time. It will be important to develop a portable standard that can be trained once, with relatively few exposures per principle, and plugged into a variety of systems: people will not want to have to retrain every new device. We will briefly discuss some possible approaches to implement this personalization (see (Malle, Scheutz, & Austerweil, 2017) for a discussion of an autonomous moral agent's desired competencies and qualities).

The most tempting approaches, given modern sensibilities, will be Deep Learning or Reinforcement Learning. Deep Learning learns complex patterns in large feature-rich datasets, and Reinforcement Learning uses reward and punishment to learn to navigate complex state-spaces. These approaches may well work for building in universal moral norms, but not for personalization: both require too much training data for individuals to provide. Some moral situations may only arise once, and it is important to either get them

right, or to correct the wrong behavior, the first time. Furthermore, current implementations of these approaches cannot explain *why* they got an answer. When the system makes a decision the human disagrees with (as it inevitably will), it will need to provide an explanation that the human finds satisfactory, or the human will quickly stop trusting the system.

Another possible approach is to use collaborative filtering, which predicts a user's behavior based on their similarities and differences with other users. However, people might not like being grouped with those with whom they have moral disagreements, even if they agree in some areas. A libertarian and a liberal socialist might agree that the government has no business regulating adults' personal sexual relationships or drug use, but disagree on publicly funded healthcare. Predicting on the basis of the first several shared features that they will share the latter will steer the system wrong. Collaborative filtering also has similar problems to Deep and Reinforcement Learning concerning volume of training data and quality of explanations.

Rule-based systems can incorporate contextual information and readily generate explanations. However, for our purposes rules would have to be learned from a small number of noisy exposures. Rules will only fire if conditions precisely match their triggers, and the system must know how rules trade off with each other (as morals do). In a moral domain with messy real-world data and an enormous number of inputs, these can be significant challenges.

Graphical models such as Bayes-Nets, which encode probabilistic conditional dependencies, may be effective at taking context into account and trading off values. However, Bayesian learning is computationally expensive, especially with high-dimensional input like real-world data, and can require large amounts of input data or a carefully crafted prior.

Finally, Case-Based Reasoning (CBR) has a long history in legal justification and reasoning, and has been investigated as a moral-reasoning technique (e.g. (Blass & Forbus, 2015)). CBR can work from single examples and use whatever information is provided in the case: reasoning is as rich as the input provided and the adaptation technique. Explanations are generated through mapping and adaptation. Experiences and instruction alike can be stored as cases with which to reason, and reasoning can be done on the basis of a partial match. The challenge with a CBR approach is that the system needs to have a relevant case to apply, and needs to know how to adapt it to the current situation.

As to what should be personalizable, a starting place may be decisions that primarily impact the proprietor of the system and their immediate social circle, such as decisions that involve disbursing the proprietor's resources (financial or otherwise). In cases such as hospital management systems and law enforcement support, the proprietors may be groups of people. Of course, anything that conflicts with the established overriding social ethical concerns must be exempted.

Embedding ethical principles into autonomous systems will be a time-consuming and error-prone process. The classic book *I, Robot* illustrates the challenges involved even in simple rules such as "Do what I say except when it harms others". Even assuming we can build the core ethical code into the system before it goes to the end user, our morals are complex, and we should not assume the system has learned our preferences before we've verified that it has. These systems should therefore be designed as apprentices to learn over a long period of time. While they are learning, they should have a less-trusted "trainee" mode, with default behaviors that are explicitly known to be mutable. During the apprenticeship, the system will not be allowed to take most actions without checking with a human first. Humans will understand that the period of time during which their machine nags them will be finite, and that errors are likely. We do not assume that a small child that has made a moral mistake is a psychopath, we correct her; similarly, we should understand that, early in the process, the computer is expected to make mistakes. We need to have a sense of the level at which a system can reason about morality, and trust it to make decisions to the same extent we would trust humans reasoning at the same level. Even after the training period is over, the system should have a confidence threshold below which it will consult a human. And explanation is crucial: if

a system makes a decision with which a user disagrees, but can provide a reasonable explanation, grounded in norms, as to why that decision was made, the user might still accept the system as being ethically competent (and might accept the decision). If the system cannot explain its decision, however, the human will rapidly lose trust in it.

This issue is pressing for several reasons. First, autonomous systems are already making moral decisions, as we have seen. Automated managers are in hospitals; automated assistants are in offices; self-driving cars are on the road; Hello Barbie is in homes. Our phones already help us violate social norms (by letting us drunk-dial); they should also be helping us uphold them. But the bigger reason this issue is of current importance is that this work needs to be done *before* these systems become truly ubiquitous. We need to collectively determine what it is we all agree on and how those common values will trade off with each other before systems are built that simply do it for us. Companies need to know what it is they must (or may) build into their systems. And it will be useful to have a common set of standards that an end-user can train once, then carry from system to system. Fundamentally we are talking about a set of laws and industry standards, and developing those takes time. That must be done now, in advance of pervasive deployment of these systems, rather than attempt to regulate them after they are in widespread use.

Conclusion: An Urgent Frontier

Let us finish with a question barely addressed here but that requires an answer to achieve the above goals: when should systems actively prevent their users from doing things that are illegal, or just wrong? This issue was touched upon in the Hello Barbie example, which points to the beginning of an answer. Certainly in some cases these systems should intervene to prevent harm: automatic braking systems, for example, can already prevent humans from running over people. But in most cases, personal devices should not act as law enforcement. Laws overlap and interact, and enforcing them would require autonomous systems to be legal scholars. People are also unlikely, for example, to buy a car that writes itself parking tickets. If these systems do not enforce laws and regulate illegal behavior, however, they will have to navigate the grey area between *allowing* something to happen and *enabling* it. Is a robot that stands still and allows its user to climb on it in order to crawl through a window participating in a break-in? Is a robotic wheelchair that takes its driver to a drug-dealer helping buy drugs?

Balancing the needs of the group against the freedom of the individual has been long been one of humanity's central projects. With advances in Artificial Intelligence, this old problem moves into new territory. Now is the time to begin navigating the tension between protecting society's interest and empowering people to have systems that reflect their personal convictions. When distinct ethical systems are equally compatible with a safe and well-functioning society, imposing one of them on someone who adheres to another goes against freedoms at the center of pluralistic societies. Whenever possible, we should leave these options open for the users of autonomous systems, while being careful not to give people the power to exploit and oppress others. The line is wide and blurry, and we will need to determine the answers to these questions soon. To simply do nothing is to force this burden upon the programmer, but this is rightfully society's burden to bear.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). *Machine bias.* Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Blass, J. A., & Forbus, K. D. (2015). Moral decision-making by analogy: Generalizations versus exemplars. In *29th aaai conference on artificial intelligence* (pp. 501–507).

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62.

Garvie, C. (2016). *The perpetual line-up: Unregulated police face recognition in america*. Georgetown Law, Center on Privacy & Technology.

Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., & Williams, B. C. (2016). Embedding ethical principles in collective decision support systems. In *Aaai* (pp. 4147–4151).

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of personality and social psychology*, *65*(4), 613.

Hristova, E., & Grinberg, M. (2016). Should moral decisions be different for human and artificial cognitive agents? In *38th conf. of the cognitive science society.*

Lee, K., Cameron, C. A., Xu, F., Fu, G., & Board, J. (1997). Chinese and canadian children's evaluations of lying and truth telling: Similarities and dfferences in the context of pro-and antisocial behaviors. *Child development*, *68*(5), 924–934.

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In *A world with robots* (pp. 3–17). Springer.

Pejsa, T., Bohus, D., Cohen, M. F., Saw, C. W., Mahoney, J., & Horvitz, E. (2014). Natural communication about uncertainties in situated interaction. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 283–290).

Sachdeva, S., Singh, P., & Medin, D. (2011). Culture and the quest for universal principles in moral reasoning. *International Journal of Psychology*, *46*(3), 161–176.

Scheutz, M., Malle, B., & Briggs, G. (2015). Towards morally sensitive action selection for autonomous social robots. In *Robot and human interactive communication (ro-man), 2015 24th ieee international symposium on* (pp. 492–497).

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321–326.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, *4*(6), 479.

Joseph Blass is a JD-PhD Candidate studying AI, Cognitive Modeling, and Law at Northwestern University. He is particularly interested in how humans can teach AIs morals and ethics, and how those systems can in turn properly make and justify their moral and ethical decisions.