Goal-Oriented Design for Ethical Machine Learning and NLP

Tyler Schnoebelen
Decoded AI
tyler@aya.yale.edu

Abstract

The argument made in this paper is that to act ethically in machine learning and NLP requires focusing on goals. NLP projects are often classificatory systems that deal with human subjects, which means that goals from people affected by the systems should be included. The paper takes as its core example a model that detects criminality, showing the problems of training data, categories, and outcomes. The paper is oriented to the kinds of critiques on power and the reproduction of inequality that are found in social theory, but it also includes concrete suggestions on how to put goal-oriented design into practice.

1 Introduction

Ethics asks us to consider how we live and how we discern right and wrong in particular circumstances. Ethicists differ on what they consider fundamental: the actor's moral character and dispositions (virtue ethics), the duties and obligations of the actor given their role (deontology), or the outcomes of the actions (consequentialism). Computational linguists do not need to answer a question of primacy, but the three themes of virtues, duties, and consequences do need to be considered.

This paper uses goals to draw out each of the three themes. Goals are states of affairs that people would like to achieve, maintain, or avoid in the face of changes and obstacles. The use of "goals" here is expansive so that it includes not just designers and users of a system, but also those who are (or would be) affected by the system.

NLP practitioners design and build technologies that connect to law, finance, education and main other domains that substantially affect people, often those with less access to resources and information. Privileged positions come with responsibilities. Namely, to recognize that systems affect people unevenly. To design with virtues, duties, and consequences in mind is to recognize the limits of one's perspective and then design systems with these limitations in mind.

2 Wicked problems

Simple NLP problems and simple NLP projects require you to identify stakeholders, articulate their goals, and build a plan. Another category of complex problems includes those that are only actually complex until they are decomposed into multiple simple problems.

A third category is wicked problems: those in which you can articulate goals but they are fundamentally in conflict (Rittel and Webber, 1973). For example, a traffic planner wants to build a highway because they want less congestion. But community members don't want their neighborhood cut in half because it destroys their goal of affiliation.

Wicked problems have no definitive solution because there are multiple valid viewpoints: you cannot take for granted that there is a single objective that will let you judge your solution as correct and finished.

We often shield ourselves from ethical problems by ignoring populations who would throw light on a project's wicked complexity. This is a good indication of an ethical problem: turning a blind eye to people who will be affected by the system but who are difficult to reach or who may have inconveniently conflicting goals.

3 An easy unethical project

In order to illustrate the ethical implications of goal-oriented design, let's take an example from machine learning that most readers will find straight-forwardly problematic. Here are two conclusions from an abstract on automated infer-

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 88–93,
Valencia, Spain, April 4th, 2017. © 2017 Association for Computational Linguistics

ence of criminality using faces (Wu and Zhang, 2016):

> All four classifiers perform consistently well and produce evidence for the validity of automated face-induced inference on criminality… Also, we find some discriminating structural features for predicting criminality, such as lip curvature, eye inner corner distance, and the so-called nose-mouth angle.

Can the goal for this project be simply stated? It seems to be, "Improve safety by having computers automatically detect the criminality of people's faces." This goal inherently categorizes people by degrees of criminality based on physical characteristics. It takes the perspective of the safety-minded, yet anyone categorized as criminal has legitimate goals to consider. Regardless of how many iterations the models go through, meeting the main goal will always create a group of criminal-looking people who will not agree with that definition and its consequences. This is a wicked problem.

In the United States, people of color have radically different experiences with the criminal justice system than white people. Attempting to use U.S. police data for training will not work: the criminal justice system in the U.S. is systemically biased, as can be seen in Hetey et al. (2016), which shows racial biases in Oakland police stops, searches, handcuffings, and arrests.

The Hetey et al. data is not merely counts of police actions on different kinds of civilians; it is also an examination of the differences in the aggressiveness of the language used by the police with African-American men. In other parts of the justice system, language—non-standard dialects—causes crucial testimony to be ignored (Rickford and King, 2016). Linguistic profiling is common in housing and many other areas (Baugh, 2016).

Ignoring the social and ideological uses of language means ignoring some of the way NLP techniques are applied. There are multiple companies working on models that use language data to decide who to give loans to. As with police stops, the features detected are not intentionally racially biased but they have the same effect in excluding specific individual from access to credit because of who they look, sound, or read like.

Such wicked problems are adjudicated by acts of authority. Neither wicked problems nor adjudication are inherently unethical. But dismissing the claims and goals of affected populations usually is. Such populations get hit by a double whammy: they are unlikely to be represented by technologists and other stakeholders and they have much less room to maneuver in whatever system is built.

4 The trouble with training data

The data for Wu and Zhang (2016) comes from China and includes only men, but it is ethically safer to assume that data from the ministry of public security and various police departments is biased than it is to assume that it is balanced and representative.

Interrogating training data is important for building effective machine learning models and it's also important for building ethical ones. Machine learning techniques depend upon training data, which causes two kinds of problems. The first problem is that whatever you build, it's biased towards the contexts that you can sample and the ways you get it annotated. Some populations are overrepresented and some are underrepresented.

The second problem with training data is that whatever your categories are, they are wrong. Categories can still be meaningful and useful, but it is a mistake to consider them to be natural or uncontestable. As Bowker and Star (1999) discuss, categorization always valorizes some point of view and erases others.

For example, gender detection is common in NLP. These projects typically begin with the assumption that a binary division of humans is relevant. But as Bamman et al. (2014) show, even binary models with high accuracy are descriptively inadequate. This is also the central point of intersectionality: people are not just the sum of different demographic characteristics (Crenshaw, 1989).

Goals for binary-gender detection projects are generally couched in terms of understanding people. But to what end and in which ways? Making goals explicit can help uncover latent biases in your mental model of what kind of people there are in the world and how you believe they move through it.

5 What you think of people

The ethics you adopt has a lot to do with what you think of human beings. In the case of Wu and Zhang (2016), tying facial structure to criminality suggests that some humans are "bad".

Plenty of serious thinkers have considered people to be fundamentally good. For example, that's what Confucian thinkers like Mencius and Wang Yang Ming believed (Chan, 2008).

Evidence suggests that individual choices—our goodness and our badness—are strongly dependent upon context. For example, what happens when you give theology students a chance to help a stranger? Darley and Batson (1973) demonstrate how localized the choices are: students in a rush to give a talk do not help—even when the talk they are hurrying to is about The Good Samaritan.

People commonly remember psychologist Walter Mischel's Marshmallow Test as proof that traits like self-control are destiny: certain kinds of children resist taking a marshmallow and they grow up to be successful. But the idea that people have durable traits is precisely not Mischel's conclusion. Rather, it is that people are fundamentally flexible: if you reframe how you think about a situation, you change how you react to it (Mischel, 2014).

People seem stable and consistent because they tend to be put in the same situations; in those situations they have the same role, and the same kinds of relationships to you. How do we keep getting into the same situations? The answer requires us to appreciate individual's agentive choices as well as to recognize the social structures that give rise to and constrain those choices. The systems we build enable, enforce, and constrain choices.

Defining goals, building models, and adjudicating conflict are clear exercises of power. But the powerful have another benefit. "Power means not *having* to act, or more accurately, the capacity to be more negligent and casual about any single performance" (Scott, 1990). Systems are not equally hospitable to all people and require some to perform acrobatics and contortions to get by.

Deontologists are the ethicists who focus on duties and obligations. As people in relative positions of power, we have an outsized impact on systems and therefore greater obligations to the people who are marginalized or victimized by them (Kamm, 2008).

6 Outcomes and reiterations

Utilitarians are consequentialist ethicists famous for focusing on the goodness of outcomes (Foot, 1967; Taurek, 1977; Parfit, 1978; Thomson, 1985). Outcomes are complicated: let's say criminal recognition worked. The odds are that it would make the world marginally safer for many people. But none of us have built a system with zero false positives. So a "working" criminal recognition system would make the lives of some innocent people who were treated as criminals much, much worse. Goal-oriented ethical design requires thinking about outcomes, with a special focus on which systems are created and maintained, and how disparate the outcomes are for the people subject to the system.

To think ethically is to think self-skeptically: "What is the worst possible way this technology could be used and how sound are my mitigation strategies?" Recently, a number of American consulting firms attempted to answer a Request for Proposals from an oil-rich country that wanted to understand social media sentiment on government projects like the building of a new stadium. But stated and elicited goals and use cases are not necessarily how something will be used or even what is actually desired.

The RFP stayed open for over a year, suggesting that consulting firms had difficulty finding NLP practitioners willing to take the stated goal at face value. It has subsequently been shown that, in fact, this project was intended to identify dissidents. The ability to identify sentiment about government projects can give a voice to people about those projects, which seems positive. But the worst-case scenario is that it can find people who are negative about the government for the government to track, regulate, discipline, and punish.

Considering the system-wide consequences of models leads us back to criminality recognition. It is one thing to identify an actual perpetrator of a crime, but to identify someone who has not committed a crime is to invite harassment from the police. Corporations could also use these models to make it hard to get a job, go into stores, or open bank accounts. In short, it could become nearly impossible for certain innocent individuals to operate within the law.

Systems shape the choices people are allowed to make and therefore they shape the people—not just the people suspected of being criminals, but everyone else, too. People who are not identified as criminals by the system may come to believe it works and that others who look bad are bad. In social theory terms, "subjects regulated by such structures are, by virtue of being subjected to them, formed, defined, and reproduced in accordance with the requirements of those structures" (Butler, 1999).

It seems handy to have something else make choices that we probably would have made anyhow. Even without any algorithms, there are more choice points in our lives than we can possibly give thoughtful consideration to. That's one reason why status quos maintain themselves: we tend to do the things we've tended to do (Bourdieu, 1977; Giddens, 1984; Butler, 1999).

The more consistent our systems are and the more rapidly they converge on consistency, the more they are likely to reiterate—and possibly exaggerate—what already exists. The actions a system takes may be small. But the ramifications may not be, as is the case with news recommendation engines operating in an already partisan context.

Routines in industry often serve to reduce anxiety. But *whose* anxiety? Each human or algorithmic choice offers the possibility of disturbing the status quo, but the vast majority of the time, they reproduce what came before. By considering the goals of people affected by the systems we build, we have a better chance of seeing how much people have to conform or contort themselves to receive benefits and avoid problems. In turn, these perspectives give us a better ability to abandon projects or reconceive them to give people ways of thwarting and hindering unethical instruments and effects of power.

7 Practical recommendations

NLP practitioners are used to thinking critically about models and algorithms. Taking an ethical stance means looking at goals just as critically, which in turn requires deeper interrogation of the training data, the categories, and the effects of the system. It also means seriously considering how the outputs of the specific system being built become inputs for other systems. But how does one do this other than "thinking harder?"

Perform a premortem (Klein, 2007). In a premortem, a team at the beginning of a project imagines the project was completed and turned out to be a complete disaster. They narrate, individually and collectively, the stories of the failures. This is a generally useful way of identifying weaknesses in design, planning, and implementation. Premortems can also be used to diagnose ethical problems. Ideally, participants approach the premortem from a place of true concern for people, but premortems can be helpful even if participants are orienting to problems of human resources, public relations, and customer service.

Ask for justifications. There are lots of things you could be doing, but why do managers and executives want to do *this*? Any of the following replies should put you on Ethical High Alert:

1. Everyone else is doing it and we have to keep up
2. No one else is doing it so we can lead the pack
3. It makes money
4. It's legal
5. It's inevitable

Projects that get these responses may be ethical, but these are terrible justifications in any event (for more on problematic justifications see Pope and Vasquez, 2016). You may get an idea because competitors are doing it and you certainly want to check on legality, but we shouldn't confuse wishes, plans, and circumstances with justifications. Even if markets and the law worked to promote ethical behavior (a big if), they will necessarily lag behind new ethical problems that computational linguists, data scientists and A.I. practitioners bring forth (Moor, 1985).

List the people affected. Which groups are specifically represented in the training data and which ones are left out? Who will use the system? Who will the system itself affect, distinguishing people immediately affected from those affected as the system outputs become inputs to other systems. How awful is it to be a false positive or a false negative? Who is most/least vulnerable to the negative effects of the system? The point of making a list is to keep technical models from becoming unmoored from human beings.

Is it a WMD? Cathy O'Neill describes the three characteristics of Weapons of Math Destruction: they are opaque to the people they affect, they affect important aspects of life (education, housing, work, justice, finance/credit), and they can do real damage.

What values are enshrined? We orient ethics around dilemmas of preventing harm, but it is also worth asking whether our systems bring about good. Which values are served and which are eschewed by a planned technology? A non-comprehensive list to consider: freedom, peace, security, dignity, respect, justice, equality.

Principles and values come into conflict—there's even an adage, "it's not a principle unless it costs you something". For example, a project centered on security may have negative implications for equality. Conflict is not to be avoided, it's to be made explicit—and most difficultly, it is to be made explicit to people affected. Sweep-

ing concerns under the rug or otherwise obfuscating them are convenient solutions, not ethical ones.

8 Conclusion

Technology does not just appear and impact society; it is the product of society, carrying with it the baggage of what has come before and usually reproducing it, discriminatory warts and all. Technology does not just appear: we make it.

But as Bruno Latour points out, "If there is one thing toward which 'making' does not lead, it is to the concept of a human actor fully in command" (Latour, 2003). At a construction site you can witness builders who may have mastery but certainly not full control: materials resist, personnel get sick, the weather won't cooperate, the planning department requires another form, the client is late with payment but fresh with a new idea.

Mastery and expertise do not imply control over objects and people; they imply practice and the ability to translate that practice into both plans and improvisations.

An important aspect of virtue ethics is practicing and developing dispositions towards moral choices (Annas, 1998). To develop habits of bravery, justice, self-control, and other virtues means practicing them. By focusing on goals, we focus on the connections between systems and people. We talk to people about their goals and their situations. We reason through surface conflicts that can be solved and discover where compromise is impossible so that we know when to reimagine our systems and when to abandon them. Done consistently, this kind of design develops habits of thinking and feeling that enable and refine our capacity to be ethical and build ethically.

It is necessary to acknowledge and address Crawford (2016)'s critique: most of the people who build technology come from privileged backgrounds, which makes it difficult for our imagination and our empathy to extend out to everyone our systems will affect.

The implication extends us beyond what is comfortable for many people and organizations: to not only to attend to issues of diversity and representation, but to go out and educate communities who will be affected so that they, too, can voice their goals and values. In other words, the practice of ethical design among NLP experts leads to greater ethical capacity—but ethics are too important to be left only to experts.

References

Julia Annas. 1998. Virtue and eudaimonism. *Social Philosophy and Policy*, 15(1):37–55.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical vari ation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

John Baugh. 2016. Linguistic Profiling and Discrimination. *The Oxford Handbook of Language and Society*:349–368.

Pierre Bourdieu. 1977. *Outline of a Theory of Practice*. Cambridge University Press, Cambridge, England.

Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting things out: classification and its consequences*. MIT Press, Cambridge, MA.

Judith Butler. 1999. *Gender Trouble*. Routledge, New York.

Wing-tsit Chan. 2008. *A source book in Chinese philosophy*. Princeton University Press.

Kate Crawford. 2016. Artificial Intelligence's White Guy Problem. *New York Times*.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and anti racist politics. *University of Chicago Legal Fo rum*:139–167.

John M. Darley and C. Daniel Batson. 1973. " From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology*, 27(1):100–108.

Philippa Foot. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15.

Anthony Giddens. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, Berkeley, CA.

Rebecca C Hetey, Benoît Monin, Amrita Maitreyi, and Jennifer L Eberhardt. 2016. Data for Change: A Statistical Analysis of Police Stops, Searches, Handcuffings, and Arrests in Oakland, Calif., 2013-2014.

Frances Myrna Kamm. 2008. *Intricate ethics: Rights, responsibilities, and permissible harm*. Oxford University Press.

Gary Klein. 2007. Performing a project premortem. *Harvard Business Review*, 85(9):18–19.

Bruno Latour. 2003. The Promises of Constructivism. In *Chasing technoscience: Matrix for materiality*. Indiana University Press, Bloomington, IN.

Walter Mischel. 2014. *The marshmallow test: understanding self-control and how to master it*. Random House.

James H. Moor. 1985. What is computer ethics? *Metaphilosophy*, 16(4):266–275.

Derek Parfit. 1978. Innumerate ethics. *Philosophy & Public Affairs*:285–301.

Kenneth S. Pope and Melba JT Vasquez. 2016. *Ethics in psychotherapy and counseling: A practical guide*. John Wiley & Sons.

John R. Rickford and Sharese King. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and be yond. *Language*, 92(4):948–988.

Horst WJ Rittel and Melvin M. Webber. 1973. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169.

James Scott. 1990. *Domination and the Arts of Resistance: Hidden Transcripts*. Yale University Press, New Haven, CT.

John M. Taurek. 1977. Should the numbers count? *Philosophy & Public Affairs*:293–316.

Judith Jarvis Thomson. 1985. The trolley problem. *The Yale Law Journal*, 94(6):1395–1415.

Xiaolin Wu and Xi Zhang. 2016. Automated Inference on Criminality using Face Images. *arXiv:1611.04135 [cs]*, November. arXiv: 1611.04135.