Chapter 14

Machine Learning and Irresponsible Inference: Morally Assessing the Training Data for Image Recognition Systems

Owen C. King
Department of Philosophy, University of Twente, The Netherlands
owen@owencking.net

https://doi.org/10.1007/978-3-030-01800-9_14

Abstract Just as humans can draw conclusions responsibly or irresponsibly, so too can computers. Machine learning systems that have been trained on data sets that include irresponsible judgments are likely to yield irresponsible predictions as outputs. In this paper I focus on a particular kind of inference a computer system might make: identification of the intentions with which a person acted on the basis of photographic evidence. Such inferences are liable to be morally objectionable, because of a way in which they are presumptuous. After elaborating this moral concern, I explore the possibility that carefully procuring the training data for image recognition systems could ensure that the systems avoid the problem. The lesson of this paper extends beyond just the particular case of image recognition systems and the challenge of responsibly identifying a person's intentions. Reflection on this particular case demonstrates the importance (as well as the difficulty) of evaluating machine learning systems and their training data from the standpoint of moral considerations that are not encompassed by ordinary assessments of predictive accuracy.

Keywords Machine learning algorithms · Image recognition systems · Training data · Responsible AI judgment · Ingrained responsibility · Modular responsibility · Intention ascription

It appears in
D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134, pp. 265-282.
https://doi.org/10.1007/978-3-030-01800-9

14.1 Introduction: Humans And Computers Drawing Conclusions Responsibly

Consider Ned, who does not know the difference between a peanut and a cashew. In fact, *cocktail nut* is the most specific category in this region of Ned's gastronomic conceptual taxonomy. Now suppose I've taken it upon myself to teach Ned to see the difference. So, I show him five labeled photos of peanuts and five labeled photos of cashews. Then I show him a new picture of a nut without a label. He confidently says, "Peanut!" and he is correct. I show him a bunch more new photos, and he identifies all the peanuts and cashews correctly. Mission accomplished. Ned has learned to visually discriminate peanuts and cashews.

Machine learning systems for image recognition operate much the same way. They are fed sets of images paired with descriptions, which are the *training data*. And then the systems generate descriptions for (or match pre-given descriptions to) new images. It amounts to an advance in image recognition when a system can draw more accurate conclusions than previous systems on the basis of the same training data. But this is not the only sort of improvement possible. Training data can be improved, too. The set of images could include more relevant variety, or the descriptions could be more accurate, or the data set could just be more voluminous. In our example of Ned, better training data might mean teaching him using sharper images of peanuts and cashews. Probably images that showed differences in the textures of the two types of nuts, all else equal, would be more helpful to him than images that lacked this level of detail.

It is tempting to think that if one set of training data yields computer systems that draw more accurate conclusions than those from systems trained on other data, then the data set that yields the more accurate systems is better. But I do not think this is the whole story. As machine learning systems, such as image recognition systems, become more and more sophisticated with wider and wider application, it is not just the accuracy of the conclusions that matters. Just as a judgment pronounced by a human might have been irresponsible, despite its accuracy, computer systems also can draw conclusions irresponsibly though accurately. And this irresponsibility can be due to the data on which the systems were trained.

Here are a couple cases of human judgment that exemplify the kind of worry I have in mind. Suppose we have an image of a man running behind a running woman who has a frightened look on her face. Suppose I look at the image and say, "He's trying to hurt her!" Well, I might very well be correct. But, clearly, my judgment has overshot my evidence. What if the man and the woman are both fleeing from some other menace? Or suppose we have a photo of a man and a woman, both finely dressed, smiling as they sit at a candle-lit table with an elegant dinner laid out before then. If I say, "They're on a date," then my judgment has gone too far again. Perhaps they're just friends; it might even be that they're both gay.

Note that the worry here is not just epistemic. We might even suppose that images that look relevantly like the first one 98% of the time really do picture one person trying to hurt another, and we might suppose that 98% of images relevantly like

the second really do depict dates. The irresponsibility involved here is more about *failure of respect* than about a lack of evidence; it is *more ethical than epistemic*.

Suppose I am barely acquainted with the two people—call them Jack and Cleo—shown in the dinner picture. And suppose I was at that restaurant and happened upon that very scene. If I ran into Jack at the coat check, I wouldn't say, "How's the date going?" Expressing the judgment that they're on a date would be presumptuous. And my embarrassment will be fitting if Jack says, "I'm gay, you idiot." But even if Jack and Cleo are indeed on a date, my comment would be no less presumptuous. The problem is that my inference was based on a superficial pattern they seemed to fit, rather than on any intentions they had expressed or any other facts about them as individuals.

Now if we have an image recognition system trained on data that include judgments like those in the two examples I just described—the example of the running people and the dinner example—then the irresponsible (though perhaps quite accurate) judgments will affect the way the system operates subsequently. Irresponsible judgments in the training data are likely to yield irresponsible conclusions at runtime.

The main issue here extends beyond just image recognition. The general issue is about the responsible use of AI systems capable of making judgments that might carry some moral weight. How ought the developers of these systems ensure that the systems judge responsibly? One option is to train the systems just for statistical accuracy, and then add an extra layer of processing to ensure that the judgments are applied responsibly. A second approach is to train the responsibility into the system from the beginning, by ensuring that the set of training data does not encode some pattern of irresponsibility. We can think of these approaches as *modular responsibility* and *ingrained responsibility*, respectively.

In the rest of this paper I will consider how we might achieve ingrained responsibility for machine learning systems, especially image recognition systems, that draw conclusions about what actions persons perform. This focus is attractive because of the present and ongoing advances in the development of such systems. In general, I suspect that it is prudent for us to prefer ingrained responsibility over modularization. But, as we will see, the temptation to modularize responsibility will be strong.

14.2 Presumptuous Judgment

Before returning to issues about image recognition and machine learning, it is worth elaborating the central moral concern here. The basic worry is that some judgments about a person's actions may be objectionably presumptuous. My goal is not to give a comprehensive account of presumptuousness or the reasons it is objectionable, but I hope to say enough about it to illuminate the sort of worry I have in mind.

We can say a judgment of a person's intentions is *presumptuous* when the intentions were ascribed on the basis of superficial features of the person, instead

of on the basis of the person's own individual profile of past and present mental states. This way of characterizing presumptuousness is not intended to be a precise definition that draws a sharp boundary around all the cases of presumptuous judgment. It is quite possible that our thinking about these issues is too hazy and mutable to make drawing a sharp boundary desirable or even feasible. Instead, what this characterization does is locate and orient presumptuous judgment with respect to types of possible evidential bases. The more a judgment of a person's intentions is based on facts about that particular individual's thoughts and desires—as manifested in, say, prior action or speech—the less presumptuous it is. The more the judgment is based on other characteristics of the person—especially general, population-wide patterns she seems to fit—the more presumptuous it is. I will not provide here a thorough defense of the claim that presumptuousness is morally problematic, but I will try to say a little bit to make the claim plausible.

First, it is worth observing that many among us (including myself) tend to be offended when people make unwarranted assumptions about our desires, goals, and intentions. Consider this scenario: Suppose I have an acquaintance, Silas, who is a bit overweight. I overhear a conversation in which Silas mentions that he has planned a trip to the beach several months from now. So I infer that Silas intends to lose weight. (He wouldn't want to look fat in his swimsuit, right?) Then, when a mutual friend is preparing for a dinner party, to which Silas and I are both invited, I suggest that she include only light fare on the menu, since (I believe) Silas is trying to lose weight. Now, as it turns out, Silas is not at all concerned about his weight. It would be fitting for Silas to be offended, or at least annoyed, at my presumptuousness. Note that the problem is not that my inference was terribly faulty from a purely epistemic standpoint; it was that I made an inference (which I then acted upon) about Silas's intentions, even though I did not know enough about Silas to do so responsibly. So, I should have withheld judgment, or at least abstained from acting on my judgment.

For another example, consider another scenario involving Silas. Suppose Silas decides to send his daughter to the local public high school instead of the nearby, expensive, private high school. Upon hearing about this, Silas's neighbor Albert infers that Silas is trying to save money. As it turns out, Silas's choice was motivated by his hope that his daughter will benefit from an education among a more inclusive group of students. Here again, it would be fitting for Silas to react with offense or annoyance at the presumptuous judgment.

Second, note the close link between presumptuousness and stereotypes. We can understand many stereotypes as constituted by shared patterns of presumptuousness. For example, imagine that Ravi is an Indian-American college student whose parents immigrated to America shortly before he was born. At college Ravi chooses pre-med as his major. Peter, Ravi's roommate, assumes that Ravi is aiming to become a doctor because of pressure from his demanding parents. It turns out that Ravi has always been interested in human biology and the practical applications of it. Peter's presumptuous judgment about Ravi's goals was a manifestation of a general stereotype Peter has accepted about Indian-Americans. Note that even if Peter had been correct about Ravi's motivations, basing his judgment on a stereotype

about Indian parents still would have been inappropriate. It is not hard to think of cases of stereotyping that are much more pernicious than this one.[1]

Finally, consider this not-too-far-fetched example, which is a bit more like the image recognition cases that are our main concern. Imagine Tara, who is an academic advisor at a large state university in the U.S. One of her duties is to have one-on-one meetings with incoming students to help them choose and register for courses during their first semester at college. Now, after a couple of years of conducting these meetings, Tara realizes that the meetings would be much more productive if she proposed a default schedule to the student at the beginning of each meeting. So, she tries this, and each student starts with a default schedule that includes Calculus I, First-year Writing and Composition, Problems of Philosophy, and Intro to the Life Sciences. At first, she does not customize the schedule for each student because she has little information on which to base any recommendations. Because of a poorly conceived information and record system at the university, she has just a photo of the student and the student's home address. However, after the first year of using her new system, and despite her dearth of background information, Tara happens to notice one regularity: Male students who hail from the northern part of the state and who are pictured in preppy attire always want to sign up for Intro to Business. So, Tara adjusts her system. For most of her advisees, she continues to offer that original default schedule. However, for her preppy, northern males, she includes the introductory business course in place of the life sciences course. After this adjustment, Tara's own personal records indicate that she has reduced her average meeting duration by 5%. So she makes the adjustment permanent.

Despite the increased efficiency from Tara's newly adjusted policy, it may strike us as suspect. But if there is a problem here, it is not inaccuracy or lack of evidence. The policy was devised on the basis of plenty of data, and it is even backed by some empirical confirmation. The problem is that she is predicting individuals' preferences (and using these predictions in ways that might influence them) on the

1 A few clarifications about the relationship between stereotypes and presumptuous judgment may be helpful. First, not all cases of presumptuous judgment involve stereotypes. Stereotypes involve associating an individual with a group (Blum 2004, Beeghly 2015). But it is possible to make a presumptuous judgment without relying on a group association. For instance, I might make a presumptuous judgment about a person's intentions just on the basis of the assumption that her goals are the same as my own. Second, not all uses of stereotypes involve presumptuous judgments. This is simply because not all stereotypes are about persons' intentions. Finally, regarding the moral features of stereotypes and presumptuous judgments: Presumptuousness, all else equal, tends to be morally undesirable, but it's controversial whether this is true of all stereotypes. Beeghly (2015) argues that not all stereotyping is morally objectionable, and Lippmann (1922) saw positive and negative aspects of stereotyping. In contrast, Blum (2004) holds that stereotyping is always morally objectionable to some degree. My contention here, that presumptuous judgments manifest inadequate respect for persons as individuals, is consistent with Beeghly's explanation of when and how stereotypes fail to respect persons as individuals. However, my thinking about why such a failure of respect is morally objectionable shares more with Blum's analysis than with Beeghly's. In the context of the present paper—with its focus on the moral evaluation of training data for machine learning systems—it is enough for my purposes if at least some judgments are morally objectionable precisely because of their presumptuousness.

basis of the persons' conformity to a superficial pattern, and thus failing to treat them as individuals. The problem is a moral one.

If indeed presumptuousness of the sort I've been gesturing at is undesirable, we will not want our computer systems to issue presumptuous judgments. As already noted, one approach, the modular approach, would have us outfit our computer systems with an additional stage of processing which took the set of statistically founded judgments and filtered out the presumptuous ones. The ingrained approach, which I'm exploring here, would effectively apply a filter on the opposite end, removing presumptuousness from the training data. To see how this would work in the case of image recognition systems, we need to look a little more closely at these systems and how they are trained.

14.3 Image Recognition And Sources Of Training Data

There are various kinds of image recognition tasks we may wish to have a computer perform. Given a photograph, we may wish to have a computer classify *what kind of scene* it is (for example, a desert or a grocery store) or identify *what objects* are pictured (for example, a camel or a cantaloupe). We might also wish to have the computer draw more nuanced conclusions—specifically about the relations among various elements and *what is happening* in the photograph (Fei-Fei, Li 2010). For example, we might like the computer to tell us that a camel is drinking from a spring or that a boy is adding a cantaloupe to his shopping cart.

Advances in computer vision in the last decade have begun to make automated scene classification and object identification more practical. And recently, new research has made headway in the third sort of task. Some new image recognition systems can tell, with some accuracy, how the objects in an image are related—reporting not just the *what*, but also the *what's going on*. This progress is the result of combining two branches of AI research: computer vision and natural language processing. The new image recognition systems integrate visual meaning and linguistic meaning in the same models, facilitating greater precision and subtlety in associating descriptions with images (Karpathy et al. 2014, Vinyals et al. 2014).

At a basic level, recent innovations notwithstanding, the new AI systems operate on the same principles as their predecessors. The first step is usually to feed the systems large sets of data. It is from this training data that a system "learns" (i.e., creates a rich model of the data). It is only once some learning has taken place that the machine learning system becomes useful. (Whether the learning process continues once the system is in operation depends on the specific system and its implementation.) In the case of image recognition, the training data includes scores of images paired with descriptions. Different data sets include different images, and the form of the descriptions may vary as well—from single-word descriptions to multi-sentence paragraphs.

What are the sources of training data for image recognition systems? It is tempting to think we have an embarrassment of riches. The Internet, from professional media outlets to social media, provides a never-ending stream of captioned images. It is Big Data *par excellence*. Consider how e-commerce websites like Amazon and eBay analyze their unceasing streams of consumer behavior data in order to train their systems to make more intelligent product recommendations. Similarly, to train our image recognition systems, one might think that we just need to point them at the streams of captioned photos that perpetually pour from the likes of Facebook, Twitter, Flickr, Instagram, Pinterest, Imgur, etc.

But a bit of reflection shows that this approach is a non-starter. After all, why do people caption images in the first place? The goal is certainly not to give plain and literal, yet comprehensive, descriptions of the contents of the photos. Instead the goal is to tell us about the things not pictured—like important background information—that make the photo interesting. If a photo shows a chemist in her lab, the caption is likely to say who she is and what she studies. It will *not* say anything like this: "A woman with goggles and a white coat lifts a glass vessel containing blue liquid." Such a caption would be useless to us; we can notice all this (and much more) from a quick glance at the photo.[2] But this is exactly the kind of caption we need paired with our image if it is to be part of our training data. The point, then, is that the training data we need for image recognition systems—unlike paradigmatic big data applications in which the relevant data sets continually accrete through the everyday course of events—must be artificially created and collected.

Artificial creation of training data is a daunting task, but it's not quite as difficult as it might initially seem. Researchers and developers can simply hire people to describe photos. And with *crowdwork services*—like Amazon's Mechanical Turk—which crowdsource the completion of large sets of microtasks, it is fast and inexpensive to create large sets of training data. Researchers can define tasks and advertise them within Mechanical Turk, and then human workers (the "Turkers") find them and complete them. In 2009, computer vision researchers at the University of Illinois used Mechanical Turk to acquire human-generated descriptions for over 8,000 images from the Flickr photo sharing website, in less than twelve days and at a cost of less than $1,000 (Rashtchian et al. 2010). The result was a data set known as Flickr 8k, which includes approximately 8,000 images paired with the descriptions written by Turkers (Hodosh et al. 2013). Thus, crowdwork takes care of the major practical obstacle in the way of training image recognition systems.[3] So, now we can begin worrying about ingrained responsibility—what it takes to make sure that none of the image labels in our training data express presumptuous judgments.

2 As Hodosh et al. (2013) point out, "Gricean maxims of relevance and quantity entail that image captions that are written for people usually provide precisely the kind of information that could not be obtained from the image itself, and thus tend to bear only a tenuous relation to what is actually depicted."

3 Though crowdwork raises ethical issues of its own (Marvit 2014).

14.4 Integrated Responsibility For Still Photographic Training Data

How could a group of workers—individuals paid to label images—produce training data that encodes responsible judgments about what people depicted in the pictures are doing? The simple answer is that the workers must adhere to strict instructions about the kind of descriptions they are to provide. If I am right that presumptuous judgments are morally objectionable, then the instructions should rule out presumptuous judgments. So, one option would be simply to instruct the workers to avoid presumptuousness.

But this sort of instruction is awfully abstract and not the most straightforward to operationalize on a case-by-case basis. Clearer instructions are required. As it turns out, Hodosh et al., the team that created the Flickr 8k data set, did an admirable job with their instructions. In a qualification test for workers who might write image descriptions, the researchers gave prospective workers this characterization of a good description:

> A good description...
> ...should provide an explicit description of prominent entities in the image.
> ...should not make unfounded assumptions about what is occurring in the image.
> ...should only talk about entities that appear in the image.[4]

The third and especially the second of these three clauses should serve to rule out many cases of presumptuousness. After all, part of what constitutes presumptuousness, as I've characterized it, is an inappropriately grounded judgment about what is motivating a person. So, my complaint is only that these instructions are not strict enough in what they prohibit. As we've seen, a judgment may be well-founded, in that it is statistically well-supported, yet presumptuous nonetheless. If presumptuousness is indeed undesirable, the rules for making assumptions about persons' actions should be more strict than the rules for making assumptions about other sorts of occurrences.[5] For example, the graphical data in an image depicting the view from a window looking out into a rainy day may be consistent with the unlikely possibility that the falling drops of water are coming from a sprinkler somewhere off to the side, but that wouldn't make the judgment that it's raining inappropriate. In order for our image recognition systems to be as useful as possible, we would prefer an image that appears to depict rain be described as depicting rain.[6] A 2% chance that it is not actually raining is not enough to withhold the

4 This comes from the online appendix to Hodosh et al. 2013.

5 This suggests another way to explain what is wrong with presumptuous judgment. To judge a person's mental states according to a standard like we would use for any other sort of judgment not involving persons, is to take what Peter Strawson (1962) called the "objective attitude" rather than the "participant attitude" toward the person.

6 Of course, the image recognition system could report the falling water, and we could rely on some other process to infer from the falling water that it must be raining. But this would be to limit

Fig. 14.1 An image from the Flick 8k data set

judgment that it is raining. However, a 2% chance of error is enough to withhold the judgment that the dining man and woman are on a date.[7] That is because there is more to avoiding presumptuousness than making judgments with sufficiently high probability.

To instruct workers in such a way that their descriptions avoid presumptuousness, I propose the following addition to the instructions used by Hodosh et al.: *Do not give a description of an action such that the person could plausibly deny that that's what she was doing.* As with the original instructions, some vagueness remains. However, the meaning of the instructions can be demonstrated with examples. (And such examples could be included with the instructions to the workers.)

Consider Fig 14.1, in which a woman and a young boy stand next to a table covered with various foodstuffs. This image, along with five English descriptions written by Turkers, is included in the Flickr 8k data set.

too much the capacities of image recognition systems. A scene can be one that *looks rainy*, and looking rainy may be both more intuitive and more useful information than the report that *it looks like water is falling from above*.

7 There's nothing special about the specific probability values of 0.02 and 0.98, besides the former being small and the latter being large. These values are just convenient for purposes of illustration. Values of 0.01 and 0.99 or 0.05 and 0.95 would have worked just as well (although values that were too extreme or too moderate would indeed alter the examples).

The Turkers' descriptions of Fig. 14.1 were as follows:

1. A woman and a boy are making hamburgers in the kitchen.
2. A woman in a white shirt prepares a large meal of hamburgers.
3. A woman is holding a jar of mustard and a boy is looking at a tray of hamburgers.
4. The woman has a blue shirt on with a kid to her side, and she is making hamburgers.
5. Woman and young boy stand in a kitchen with a spread of burgers in front of them.

Among these five descriptions, only (3) and (5) would be acceptable according to the additional instruction I am proposing.[8] The others make presumptuous inferences about the woman's intentions. It is clear that the woman is holding a mustard jar and sticking some kind of utensil in another jar on the table. However, it is unclear what she intends to be doing. She might be just taking a hamburger for herself; or perhaps she is just sampling the mustard. (Returning to the point I noted earlier about the relationship between presumptuousness and stereotypes, it is worth wondering whether the Turkers would have written different descriptions if the picture had included an old man in a suit instead of a young woman in a casual blouse!)

Figure 14.2 is another image from the Flickr 8k data set.

The descriptions of Fig. 14.2 were as follows:

1. Four people are lining up to purchase tickets at the theater.
2. Four people standing outside of an outdoor ticket booth.
3. Four people wait outside in a line for ticket.
4. The man and woman at the window are turned around to the man and woman behind them.
5. Two men and two women standing at the window of a ticket booth.

Among these descriptions, (1) and (3) would be prohibited by the rule I have proposed. The reason is that they attribute intentions to the persons depicted. We are not in a position to know that these people are indeed trying to acquire tickets. They might be there just to ask a question, or perhaps they are in line to get a refund, not make a purchase at all.

Despite these examples of how the instruction I've proposed would have affected this data set, I must point out that the change would be very minor. If the workers writing descriptions had adhered to my proposed instruction, the Flickr 8k data set would *not* be very different than it is. That is because the workers attributed intentions to the individuals pictured fairly seldom. This is good news. It means that the data set is useful for training image recognition systems, without much risk of generating presumptuousness.

But now we are in a position to observe that this success comes with a cost. The fairly strict limit on what is allowed in the descriptions limits the scope of

8 I do not intend this as a criticism of the Flickr 8k data set. Violations of the instruction I am recommending seem to appear only rarely in the data set. However, this image and the next are valuable for illustrating the worry I that is my focus.

Fig. 14.2 Another image from the Flick 8k data set

the judgments that can be produced by a system trained on such a data set. The descriptions of the activities in the training images are to be limited to, at most, the *overt behavior* of the persons pictured. So, the captions can describe *intentional actions* in only a very thin sense. For instance, we might say of a photo that it shows a woman kicking a soccer ball, but we cannot say that she is passing or shooting—at least not on the basis of a single still image. Necessarily missing is any attribution of aims, attempts, plans, or processes. And if the training data lacks these sorts of attributions, then a system trained on these data cannot possibly attribute them either. If we want a machine learning system to provide rich, informative descriptions of intentional actions, but also do so in a non-presumptuous way, then we will have to broaden the training data.

14.5 Theoretical Grounds For Ascribing Intentions?

We have seen that if we adhere strictly to the sort of principle I've advanced, descriptions generated by a system trained on data like the Flickr 8k data set will be limited in their informativeness. Such a system can offer very little in the way of responsible judgments about persons' intentional actions. However, many applications—indeed any applications designed to intelligently assist a user with the achievement of her goals—will need information about the user's intentions.

It is tempting here to fall back on a general idea about the basic conditions of successfully interpreting—making sense of the thoughts, behavior, and speech of—one another. Let me explain. W. V. Quine famously argued that radical translation—the process of translating the previously unknown language of a foreign speaker into one's own language—requires applying a principle of *rational accommodation*, what's more commonly known as a *principle of charity* (Quine 1960). The principle is required when trying to make headway in a situation in which the only evidence available to an interpreter is the overt behavior (including utterances) of the foreign speaker whose language the interpreter is trying to understand. The behavioral evidence will necessarily be compatible with many different translations, given the many different background beliefs the foreign speaker may hold. In such a situation, making any headway requires the interpreter to assume that many of her beliefs agree with those of the foreign speaker. So, perhaps we need to do something similar in attributing desires and intentions?

Along these lines, Donald Davidson argued that a principle of charity should be extended to the posits about what desires or values a person has. Davidson (2004b) explains the enlargement of the scope of the principle of charity this way:

> For in the plainest cases we can do no better than to interpret a sentence that a person is selectively caused to hold true by the presence of rain as meaning that it is raining... It follows that in the plainest and simple matters good interpretation will generally put interpreter and interpreted in agreement... Just as in coming to the best understanding I can of your beliefs I must find you coherent and correct, so I must also match up your values with mine; not, of course, in all matters, but in enough to give point to our differences. This is not, I must stress, to pretend or assume we agree. Rather, since the objects of your beliefs and values are what cause them, the only way for me to determine what those objects are is to identify objects common to us both, and take what you are caused to think and want as basically similar to what I am caused to think and want by the same objects.

This may seem to justify some leeway for workers writing descriptions to ascribe intentions to an agent depicted in some image, even when the image is consistent with several alternative claims about the agent's intentions. Perhaps we have no other way forward. But I do not think this is so. It is far from clear that this sort of charity is appropriate when the interpretive activity is not *radical* interpretation. The principle of charity is crucial when we have yet to establish that we are even talking about the same objects as the person we're interpreting. However, the principle is no longer required if enough linguistic commonality has been established that the interpreter is in a position to know the meanings of the person's sentences (or if the interpreter were in a position to ask the person for clarification). Hence, though the

kind of charity Davidson describes may be a condition of interpreting others in some unusual contexts human interaction, that does not justify allowing it as a heuristic in the generation of training data for machine learning systems. After all, relying on such a principle of charity yields presumptuous judgments. It is not exactly the same sort of presumptuousness featured in the preceding examples, but it may be just as bad. Instead of supplementing the information available with inferences based on group membership (as with stereotypes), according to the present strategy, the auxiliary information would be drawn from the inventory of mental states of the person writing the descriptions. It is no less objectionable to simply assume a person's motivations are like one's own than to assume that the person is motivated like people to whom she bears a superficial similarity.

An alternative way of attempting to resolve uncertainty about an agent's intentions is not to assume that her intentions match the interpreter's, but rather to assume that her intentions align with those that are most prevalent in the population. Daniel Dennett (1989b), working very much in the same vein as Davidson, discusses how we attribute desires when we take the so-called *intentional stance* toward an entity:

> How do we attribute the desires (preferences, goals, interests) on whose basis we will shape the list of beliefs? We attribute the desires the system *ought to have*. That is the fundamental rule. It dictates, on a first pass, that we attribute the familiar list of highest, or most basic, desires to people: survival, absence of pain, food, comfort, procreation, entertainment. Citing any one of these desires typically terminates the "Why?" game of reason giving. One is not supposed to need an ulterior motive for desiring comfort or pleasure or the prolongation of one's existence.

If indeed there are desires or intentions that are shared by all persons, then it cannot be presumptuous to judge of a particular individual that she has these desires or intentions. But notice that there is a difference between ascribing a standing desire to a person and judging that the satisfaction of that desire was the intention driving a particular action. My intention when washing the dishes is more accurately described as "getting the dishes clean" or "keeping the kitchen tidy" than in terms of any of the more basic desires Dennett mentions. So, to assume of a person that all her actions are to be interpreted as intending to satisfy these basic desires is another kind of presumptuousness.[9]

Even if an overly broad appeal to basic desires is just another form presumptuousness can take, it is worth mentioning because of the distinctive worries it raises. Some aims and values may be shared across all of humanity, but many are not. The variety among our aims is a source of richness in the human experience. To attempt to limit our interpretations of an individual's intentions to a fixed set of common goals is to underestimate the diversity of human motivations. Hence it is to view the person not as an individual with a distinctive orientation to the world, but as

9 I do not mean to imply that Dennett himself is guilty of making this assumption.

an indistinctive node in a homogenous system.[10] Such a view, if regularly invoked, may result in the assumption of shared intentions becoming a sort of self-fulfilling prophecy, ultimately narrowing, rather than enlarging, the courses of action open to us.

Hence, it seems that neither imputing the intentions of the interpreter, nor imputing the intentions that are common, is an acceptable way to address the lack of information we have about what motivations drive the actions depicted in a photograph. The more general principle we may draw from this is that *information about one person's goals and intentions ought not be used to reach conclusions about those of another*. Again, the point here is ethical, not epistemic. It is the upshot of the preceding discussions of presumptuousness.

One possible response to this might be to argue that presumptuousness itself is not a problem. Perhaps presumptuousness is undesirable only when the intentions presumptuously ascribed appear immoral, embarrassing, or otherwise unattractive. Along these lines, suppose that, while shopping at the grocery store, I choose the expensive, environmentally friendly cleaning spray. My actual motive might be to avoid allergic reaction to a chemical in the standard variety of cleaning spray. But if people believe that my intention is to be an environmentally responsible consumer, I may not mind their inference too much. This suggests that we may not need to have image labelers withhold judgment about any and all intentions the agent may have, just the unattractive ones. The intuitive thought in the vicinity would be something like this: *It's okay to guess at persons' intentions, as long as we give the people the benefit of the doubt*. But this is not acceptable either. Although it may be a good rule of thumb for everyday social life, and although it may avoid some negative consequences of presumptuousness, it would be a totally inappropriate policy for our image recognition systems. It would bias the training data set in a way that would reduce its accuracy. After all, people often do have unattractive motives. To train the system as though this were not true would be to introduce systematic error into the system. We would be avoiding the moral problems at the expense of adding new epistemic problems.[11]

14.6 Going Beyond Still Photographic Data To Ascertain Intentions Responsibly

We have seen that attributing intentions on the basis of some kind of interpretive charity—whether the interpreter ascribes to the person the interpreter's own intentions, intentions that are common in the population, or intentions that paint the person in a positive light—is unacceptable. Supposing we still wish to develop

10 Cf. Blum (2004).

11 And, of course, a further worry about this strategy concerns the thorny issue about how we might go about categorizing intentions as attractive or unattractive in the first place.

systems capable of making intelligent inferences about an person's intentions, we need additional sources of data.

So, let's consider what additional data would allow responsible judgments about intentions. One limitation of the Flickr 8k data set has nothing to do with any restrictions on the descriptions the image labelers were allowed to provide. Rather the limitation is due to how the images in Flickr 8k were acquired. The researchers note that images were "manually selected to depict a variety of scenes and situations" (Hodosh et al. 2013). In effect, this means that in very few cases is any person depicted in more than one image. This fact, combined with the inherent limitations of still images, entails that the data set contains almost no diachronic information. That means that even the evidence of an agent's overt behavior is severely limited. In contrast, with several successive, timestamped photos, or with a few seconds of video, instead of just a single still image, we may have a representation of behavior sufficiently rich to ascertain more—at least something beyond the bare minimum—about an agent's intentions in acting. For instance, regarding a woman kicking a soccer ball, we might be able to say whether she is passing, shooting, or just clearing it. Unfortunately, with just a single still image we have information that is consistent with too many different possible intentions on the part of the person pictured.

Consider Fig. 14.3, which is another image from Flickr 8k.

Here are the descriptions that the Turkers wrote to describe it:

1. A man dressed for cold weather plays with a stick with his black and brown dog.
2. A man in a brown vest and glasses plays with a brown dog.
3. A man in orange pants and brown vest is playing tug-of-war with a dog.

Fig. 14.3 Another image from Flickr 8k. Note the ambiguity of the aims of the man holding the stick

4. A man tries to take a stick away from a brown dog.
5. A man tugging on a stick that a little dog has in his mouth.

All of these descriptions—except, perhaps, for (5)—display some degree of presumptuousness. Also, it is interesting to observe at least some apparent disagreement among them. While (1), (2), and (3) suggest that the man's intention is to play with the dog, (4) suggests that the man's aim is simply to get the stick. But, most importantly for present purposes, note that it would not take much additional data about this scene to make it pretty obvious which of these somewhat divergent interpretations is most correct. A few seconds of video of the scene, or a series of several photographs taken over the course of a few seconds, would likely be enough. Or, if we had a record of the man expressing a desire to play with his dog, or, alternatively, a record of him saying he intended to train his unruly canine, this might be even more helpful. This points the way to a positive recommendation, though perhaps an obvious one: *Attribution of intentions to a person, in a way that is informative, accurate, and not presumptuous, requires several data points about that particular person. Likely, the more (and the more diverse), the better.*

The task of generating training data sets that are informative, accurate, and that encode genuinely responsible judgments about persons' actions, may require using not just annotated visual information about the persons, but also data of other sorts, such as the persons' histories of verbal communication. Of course, drawing on richer data sets requires more sophisticated machine learning systems.[12] And, even with additional data about an individual, the data available may still be compatible with several different hypotheses about the person's intentions. Continuing to add more and more data about the person is the only non-presumptuous path to narrowing the set of interpretive hypotheses about a person's intentions down to just one.[13] Thus, there does, after all, appear to be a route forward that avoids presumptuousness, but it is a formidable one.

14.7 Modular Responsibility Reconsidered?

I have been considering what it would take to produce machine learning systems capable of issuing responsible judgments about the intentions with which a person acted. The approach I have considered is what I described at the outset as *ingrained*

12 Such work is already underway. See, e.g., Park et al. (unpublished ms.).

13 Along these lines, Dennett argues, "the class of indistinguishably satisfactory models of the formal system embodied in [the] internal states [of an entity toward which we might take the intentional stance] gets smaller and smaller as we add such complexities [such as a wider range of behaviors]; the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which a unique semantic interpretation is practically (but never in principle) dictated" (1989b). Notoriously, according to both Quine and Davidson, some indeterminacy may be ineliminable. However, along with Dennett, I doubt that any remaining indeterminacy poses any practical or ethical problems in the context of machine learning systems. For discussion of indeterminacy and its (in)significance, see Davidson (1984b).

responsibility. The thought was that we could create systems that issued only responsible judgments, by ensuring that the data on which these systems were trained included only responsible judgments. But, as we've seen, this approach will be difficult and so may require postponing benefits otherwise soon achievable.

Also at the outset I mentioned a *modular* approach as an alternative to ingrained responsibility. The idea would be to accept training data that embodies the problems, i.e., presumptuousness, that I have been discussing. And then the task would be to add an extra stage of processing that would prevent the irresponsibility from being propagated into applications. But note that an effective module for these purposes would not be just a simple filter. An algorithm that could accurately classify as presumptuous or non-presumptuous descriptions of actions may itself require machine learning. If that is so, then it seems better to avoid presumptuousness from the beginning, in the training data that might originally introduce it. In other words, it seems better to opt for ingrained responsibility.

A final worry about the modular approach is that, in practice, it may be tempting (for convenience or other reasons) to omit the extra stage of processing. The "responsibility module" might simply be left out by a developer who didn't consider it important enough to bother with. But then irresponsible judgments would make their way into computer systems we use, and we would likely never know.[14] For this reason also I hold out hope for a tractable approach to ingrained responsibility.

In light of this discussion of machine learning systems for image recognition, I venture that there is a more general—though perhaps unsurprising—lesson to be learned here: We ought to include moral criteria among the requirements for our machine learning systems and the data on which we train them, even though doing so poses distinctive and difficult challenges.

Acknowledgments I am grateful to Andréa Atkins and to attendees of the IACAP 2016 for discussion of these issues. A preliminary exposition of some of the ideas and arguments presented in this chapter appeared in a short essay posted on the website of the Loyola Center for Digital Ethics and Policy (http://www.digitalethics.org/).

References

Beeghly, Erin. 2015. What is a Stereotype? What is Stereotyping?. *Hypatia* 30(4): 675-691.
Blum, Lawrence. 2004. Stereotypes and stereotyping: A moral analysis. *Philosophical papers* 33(3): 251-289.
Davidson, Donald. 1984a. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.
Davidson, Donald. 1984b. Belief and the basis of meaning. Reprinted in Davidson (1984a): 141-154.
Davidson, Donald. 2004a. *Problems of Rationality*. Oxford: Clarendon Press.
Davidson, Donald. 2004b. Expressing evaluations. Reprinted in Davidson (2004a): 19-37.
Dennett, Daniel. 1989a. *The intentional stance*. Cambridge, MA: MIT Press.

14 This is a specific version of the type of problem James Moor (1985) has famously called "invisibility."

Dennett, Daniel. 1989b. True believers. Reprinted in Dennet (1989a): 13-35.
Fei-Fei, Li, and Li, Li-Jia. (2010). What, where and who? telling the story of an image by activity classification, scene recognition and object categorization. In *Computer Vision*, Cipolla et. al eds., 157-171. Berlin: Springer.
Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47: 853-899.
Lippmann, Walter. 1922. *Public opinion*. New York: MacMillan.
Karpathy, Andrej, and Fei-Fei, Li. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.
Marvit, Moshe. 2014. How Crowdworkers Became the Ghosts in the Digital Machine. *The Nation*. http://www.thenation.com/article/how-crowdworkers-became-ghosts-digital-machine/. Accessed 11 January 2016.
Moor, James. 1985. What is computer ethics?. *Metaphilosophy* 16(4): 266-275.
Park, Eunbyung, Han, Xufeng, Berg, Tamara, and Berg, Alexander. (unpublished ms.). Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition. http://www.cs.unc.edu/~eunbyung/papers/wacv2016_combining.pdf. Accessed 11 January 2016.
Strawson, Peter. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1-25.
Quine, W.V. 1960. *Word and object*. Cambridge, MA: MIT press.
Rashtchian, Cyrus, Young, Peter, Hodosh, Micah, and Hockenmaier, Julia. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 139-147. Association for Computational Linguistics.
Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.