Ethics and Information Technology
https://doi.org/10.1007/s10676-019-09512-3

ORIGINAL PAPER

Presumptuous aim attribution, conformity, and the ethics of artificial social cognition

Owen C. King[1]

Abstract

Imagine you are casually browsing an online bookstore, looking for an interesting novel. Suppose the store predicts you will want to buy a particular novel: the one most chosen by people of your same age, gender, location, and occupational status. The store recommends the book, it appeals to you, and so you choose it. Central to this scenario is an automated prediction of what you desire. This article raises moral concerns about such predictions. More generally, this article examines the *ethics of artificial social cognition*—the ethical dimensions of attribution of mental states to humans by artificial systems. The focus is *presumptuous* aim attributions, which are defined here as aim attributions based crucially on the premise that the person in question will have aims like superficially similar people. Several everyday examples demonstrate that this sort of presumptuousness is already a familiar moral concern. The scope of this moral concern is extended by new technologies. In particular, recommender systems based on collaborative filtering are now commonly used to automatically recommend products and information to humans. Examination of these systems demonstrates that they naturally attribute aims presumptuously. This article presents two reservations about the widespread adoption of such systems. First, the severity of our antecedent moral concern about presumptuousness increases when aim attribution processes are automated and accelerated. Second, a foreseeable consequence of reliance on these systems is an unwarranted inducement of interpersonal conformity.

Keywords Aim attribution · Social cognition · Artificial intelligence · Presumptuousness · Recommendation · Recommender system · Collaborative filtering · Individuality · Conformity · Ethics

Introduction: an ethical issue for artificial social cognition

How our aims sharpen, how they evolve from general to more specific, can be morally significant. According to an influential individualistic ethos, all else equal, it is better for a person's aims to develop according to her individual peculiarities than to be guided to conform with common patterns (Mill 1859/2003). A core contention of this article is that it matters not just how a person's aims develop, but also how we predict and attribute aims to a person. Basing predictions and attributions on the premise that this person will follow the path of people she resembles can be objectionably presumptuous.

As we will observe, presumptuousness is a familiar and relatively minor issue in ordinary social contexts. However, this article will demonstrate that new technologies extend the scope and moral significance of presumptuous aim attribution. This is so, in particular, with artificial recommender systems, which automatically infer and attribute aims to humans. Many of these systems are based on techniques known as *collaborative filtering*, which, by design, predict a person's preferences, desires, and aims, based on her similarities with others. In relying on the assumption that individuals will conform to common interpersonal patterns, these systems are inherently presumptuous. This raises two moral concerns about these systems. First, a foreseeable consequence of reliance on such systems is an unwarranted, yet self-perpetuating, inducement of conformity. Second, not only do these systems constitute a technological realization of an extant moral concern; they exacerbate it by accelerating it and making it more widespread, beyond everyday contexts governed by familiar social norms.

✉ Owen C. King
o.c.king@utwente.nl

[1] Department of Philosophy, University of Twente, Cubicus C319, Drienerlolaan 5, 7522 NB Enschede, The Netherlands

Published online: 23 September 2019

To raise moral concerns is not to conclude that the object of concern is undesirable overall. Even if I am correct in claiming that the recommender systems in question have morally problematic features, it is undeniable that they also answer very real human needs. For instance, they may benefit us by helping us navigate the bewilderingly wide arrays of diverse options among which we commonly must choose (Iyengar and Lepper 2000). Assessing whether and when the benefits of such systems outweigh the drawbacks requires articulation of suitable evaluative concepts. To this end, this article gives an account of presumptuousness, situating it in relation to the themes of individuality and conformity, and demonstrates the applicability of this evaluative concept to a class of increasingly common automated systems.

Before laying out the relevant themes and arguments, let us situate this discussion in relation to the fields it touches: psychology, ethics, and computing. When psychologists speak of *social cognition*, they refer to the processes by which we humans come to understand one another and ourselves (Fiske and Taylor 2016). Central questions about social cognition are about how we recognize or infer what others are thinking. Hence, social cognition includes prediction and attribution of aims. Psychologists are, of course, primarily concerned with how social cognition *actually* works in humans. However, we can also consider how it *should* work, in light of moral reasons for and against various possible social cognitive processes. This is a distinction between the psychology of social cognition and the ethics of social cognition. This distinction is already evident in discussions of stereotypes: Psychologists study how stereotypes work (e.g., Hamilton et al. 1994), and ethicists assess the moral status of stereotyping (e.g., Blum 2004). The present article examines the moral status of the social cognitive processes of aim attribution. Since we will be especially concerned with how artificial systems perform aim attribution, our topic falls within the *ethics of artificial social cognition*.

The article proceeds as follows. The next section describes an example to bring the relevant range of moral questions into view. To sharpen the ensuing discussion, the two subsequent sections define and elaborate two categories of aim attribution: over-specific aim attribution and presumptuous aim attribution. The following section invokes these categories to critique an increasingly prominent class of information technology. That section includes a high-level introduction to recommender systems and collaborative filtering, as well as an argument that such systems involve a mechanism well-suited to induce interpersonal conformity. The final section offers a broader assessment of the moral significance of presumptuous aim attribution, demonstrating why we should be concerned about its proliferation through automation.

Recognizing a moral dimension of aim attribution

Our lives unfold and take shape as our general aims evolve to become more specific. How, and in response to what pressures, this evolution occurs sometimes matters to us. To focus our attention on several morally salient features of this process, we begin with an example.

Imagine Celina who will soon begin her first year of university study. Celina would like a part-time job while she is a student, and she applies for a program that matches students with flexible jobs at her university. Her application is accepted, and so she will be placed in one of three available roles: (1) selling university branded merchandise (like t-shirts, hats, mugs, etc.) at university events, (2) monitoring parking lots to allow only permitted vehicles, and (3) assisting with clerical tasks in the Philosophy Department. Suppose that Celina will end up taking the first position, selling university merchandise. Consider three alternative ways this might come about.

First, imagine Celina receives a letter from the program administrator describing the three jobs and offering a choice among them. So, Celina consults several close friends who report enjoying the sales job. She also recognizes that this position will allow her to visit a variety of university events, which seems both prudent and exciting. In contrast, monitoring parking lots seems dull and monotonous. Also, she has heard that philosophers are insufferable. Upon reflection, she develops a desire to work in the sales position, and she responds to the letter accordingly.

Second, suppose Celina is traveling with friends when the letter arrives to her parents' home. Her parents notice the envelope is marked as containing time-sensitive information. Since Celina is far away and her parents are unable to contact her, they decide to open the letter and act on it. However, they do not know if Celina had one particular job in mind. Wondering which position Celina would want, they consider a few factors. For one thing, they know Celina thrives with exposure to new things (like she might encounter at events where she would sell merchandise). They also know Celina has often felt uncomfortable when called upon to be assertive, and they have noticed that many parking monitors are rather gruff. Finally, they have heard that philosophers are insufferable. So, they conclude that Celina would pick the sales job, and they respond accordingly on her behalf.

Third, suppose the administrator of the student employment program decides to streamline the placement process by matching applicants to positions himself (without sending out letters). For Celina, he immediately rules out the parking position because he believes that most women are insufficiently assertive. Looking at the photo in Celina's profile, he perceives Celina as cheerful and attractive. As a cheerful person, she might be able to deal well enough with the insufferable philosophers. However, he feels sure that this cheerful, attractive young woman will enjoy the attention from attendees of university events. So he concludes that Celina would want the sales position.

Celina had to undergo some change of mind, from the general aim of taking a student job, to a specific intention to take one of the particular positions. Although the outcome of the process was the same in the three alternative scenarios,[1] the scenarios do not seem to be morally on a par. The first two scenarios seem fairly unobjectionable. The first manifests a deliberate exercise of agency, with Celina responsive to relevant facts about herself and her situation. In the second, Celina's parents made the same determination that Celina would have made, and did so in ways that reflected their understanding of her as an individual, taking into account some of the same facts about her personality that would have been the basis for her own determination. Hence, especially since Celina could not have responded herself, there seems little cause for complaint regarding the second scenario.

In contrast, the third scenario is likely to raise qualms. Like the second scenario, the third is one that called for deciding on Celina's behalf, and the eventual decision was the same. However, unlike Celina's parents who drew on knowledge specifically about Celina, the administrator reasoned according to some dubious generalizations about gender. Furthermore, even putting aside whether the generalizations were accurate, we may worry about whether a generalization like that was appropriate for the case at hand. An interpersonal generalization based on Celina's superficial characteristics figured crucially in a decision that seemed to call for sensitivity to individual differences. Such a generalization exhibits what I will call presumptuousness, as I will elaborate momentarily.

At any rate, *none* of the three scenarios seems particularly unusual or unfamiliar. Moreover, even if I am correct in suggesting something is morally amiss with the third scenario, there may seem little cause for concern. After all, the administrator ended up making the right decision. No harm, no foul, we might think. To protest an acceptable answer wrongly reached may seem fussy and overly fastidious. However, withholding protest is risky. If it were increasingly common for our aims to take shape as in the third scenario, then we should worry about this ipso facto loss of individual agency and the diminished authorship of our own lives. Shortly, I will argue that just such a worry is warranted regarding automated classification systems applied to us and our aims. Before that, we must look more carefully at the specificity of our aims, the accuracy of aim attributions, and the kinds of evidence on which such attributions may be grounded.

The specificity of our aims and aim attributions

Consider how our general aims may sharpen and become more specific. Imagine I am browsing the news on a summer Saturday morning. The sun shines, and a warm breeze gently rustles my curtains. As I finish a cup of coffee, I get restless, and I am becoming inclined to be active outdoors. A friend sends a message asking, "Want to go water skiing this afternoon?" to which I immediately reply, "Yes, I do!" I have never been water skiing before, and I start anticipating what it will be like. As I read about it and watch a video, I get excited. My friend says she will be in touch again shortly, and so I return to reading the news. In a half hour, I get another message from the same friend: "Hey, want to go for a hike?" Now surprised and a little concerned, I reply, "Wait, what about water skiing?" She says, "Well, I just went skiing last weekend. And I know you like hiking." A little disappointed, I reply, "I mean, I guess I could go for a hike."

In this scenario, my aims evolved from very general and open-ended to much more specific—specific enough, in fact, to preempt other specific aims that would have satisfied my original general aim. Before my friend's first message, motivated by restlessness, I had a very general aim of some activity in the sun and the breeze. If she had first asked if I wanted to hike, my reply would have been just like my reaction to water skiing: "Yes, I do!" But, as it happened, she first suggested skiing, and I soon wanted to do precisely that.

This mundane example exhibits two facts that will be essential in what follows: one fact about a kind of diversity in our aims, and one fact about the mutability of our aims. First, our aims are diverse with regard to their levels of specificity. In other words, there is a difference between our general dispositions for pursuit and our well-articulated objectives. We can think of these mental states on a continuum from the very general to the very specific. At one end, our aims barely deserve that label; they are more like

[1] Since the story is fictional, we can stipulate that there are no lasting differences among the scenarios. For example, we can assume that any extra feeling of agency that Celina might have experienced in the first scenario, from having made the selection herself, quickly becomes negligible.

dispositions to respond favorably to particular sorts of representations and unfavorably to others. Our general, unarticulated aims are among the sorts of mental states that are most likely shared by some non-human animals. My old dog Petey was disposed to pursue a vaguely defined set of objects, including familiar subcategories, like rabbits and people food, as well as untold objects he never encountered or imagined. Such dispositions to pursue open-ended classes of objects—broad *inclinations*, we might consider them—are aims of the most general sort. At the other end of the continuum, we have highly refined, conceptually complex aims. Petey, loving though he was, never had a desire that I recoup my disastrous financial loss and regain my self-esteem before traveling home for a reunion with old friends. Dogs simply do not have the mental resources to formulate aims so specific.

Despite the diversity of aims, the category is united by a common characteristic: Aims are dispositions for pursuit. They are mental states that motivate interaction with the world, to change it in a particular direction. That is, they are mental states the world is brought to fit, in contrast to states like belief which should come to fit the world (Anscombe 1957). It is this unifying feature of aims that make them relevant to the practical topic of this paper. Aims encompass the full class of mental states that motivate the pursuit or acceptance of an option that has been suggested or offered, whether the suggestion comes from a friend or an automated system.

This understanding of aims shares much with the account of desires offered by Smith (1994). However, aims are a more appropriate focus than desires. Although any desire counts as an aim, aims are not necessarily desires, since desire is essentially connected to the reward system and pleasure (Schroeder 2004). For present purposes, we need not require that aims bear such connections, and we thus allow the possibility that a person could be guided by an aim that would not, strictly speaking, count as a desire.

Thus conceived, aims are similar to the construct of *goals* in psychology, particularly as in the goal-setting theory of motivation (Locke and Latham 1990). However, the goal-setting literature tends to focus on action and accomplishment of goals (Locke and Latham 2013), whereas aims, as understood here, also include more passive, dispositional preferences. Moreover, though consistent with goal-setting theory, the conception of aims here does not endorse or rely on the particular commitments of goal-setting theory, or its unitary concept of motivation (Deci 1992). In short, the general conception of aims adopted here is intended to be ecumenical.

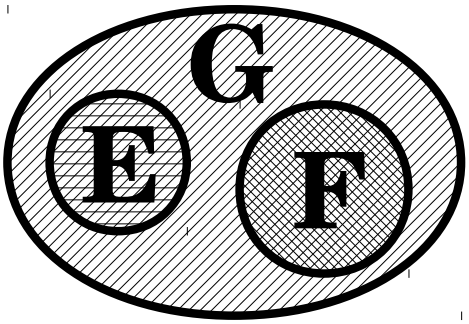

Fig. 1 G is the content of a general aim. E and F are, respectively, the contents of two mutually exclusive aims that are ways of making the general aim of G more specific. Realizing either E or F would suffice for the realization of G

The second essential fact about our aims is that they are dynamic, with general, open-ended aims evolving into, or being replaced by, more specific aims. We often aim at some general end, and, in the course of reflecting on it and pursuing it, come to aim at some more specific end that constitutes one way that the general end might be realized.[2] Indeed, this is arguably one of the main operations of our practical reasoning (Richardson 1997; Bratman 1989). But, besides orderly reasoning, this process of increasing aim specificity can happen more haphazardly as well. As in the example above, a friend might suggest a Saturday activity. Or I might see an advertisement as I browse the news. Or I might simply find myself in a setting where I am expected to behave in some way which, though indeed one among many possible refinements of a general aim, is more specific than I had yet imagined: "Oh, I guess we're heading to the dance floor. Okay, cool."

With the class of aims thus circumscribed, we can consider the attribution of aims. I will understand *attribution* broadly, such that judging that a person has a particular aim, stating that she has that aim, drawing inferences from the premise that she has that aim, and treating the person as though she has that aim, would each count as attributing the aim. And now we can consider how attributions may succeed or fail to accurately reflect the specificity of the aims a person has.

Let E, F, and G be propositions such that E entails G, and F entails G, and E and F are mutually inconsistent. We can think of G as the content of a general aim and E and F as two incompatible ways of making G more specific. For a concrete interpretation along the lines of our first example, think of G as *I engage in outdoor recreation*, E as *I go water skiing*, and F as *I go for a hike*. Thus, if a person aims that G, and it happens that either E or F comes to be, then that aim is *ipso facto* satisfied. It is possible for a person with the

[2] Of course, the reverse is always possible. A specific aim, e.g., *write a coming-of-age novel*, can evolve to something more general, e.g., *do some creative writing*. Though such specific-to-general transformations are interesting, too, they will not be in focus here.

general aim that G to aim for the disjunction of E and F, but not for their conjunction (Fig. 1).

We can distinguish three different kinds of inaccuracy in the attribution of aims. First, suppose that a person aims that E, but we attribute to her the aim that F. Then we have simply attributed the wrong aim to her. So, if it becomes true that F, then her aim will have been thwarted. Call this class of inaccuracy *simple inaccuracy*.

Next, suppose that the person aims that G, and we attribute the aim that F. In that case we have again attributed the wrong aim, but the mistake is much less severe from a practical perspective. That is because if it becomes true that F, then her aim will have been satisfied, not thwarted at all. So, we must distinguish this class of inaccuracy from the first; we can say that such attributions are *inaccurate because over-specific*. These attributions are inaccurate precisely in the sense that they represent the class of desired outcomes as *narrower* than the class that would actually satisfy the aim.

Finally, suppose that the person aims that F and we attribute to her an aim that G. In one respect, this is not inaccurate at all, since if she aims at F, she cannot deny that she aims at G (assuming the entailment of G by F is transparent). However, attributing an aim that G without also attributing a specific aim that F would mistake the content of her aim. Furthermore, such inaccuracy is of practical import since it may become true that G without it becoming true that F. Indeed, if it becomes true that G because E is true, the person's aim would be thwarted. We can say that our attribution in this case was *inaccurate because under-specific*.

Our primary focus here will be inaccuracy due to *over*-specificity. This type of inaccuracy will be especially significant for two reasons. First, this is a form of inaccuracy that may easily go unnoticed, because people tend to accept over-specific aim attributions. Second, over-specific attributions, despite their inaccuracy, influence and direct behavior.

Our earlier example shows how people tend to accept over-specific aim attributions. Suppose, as in the example, by mid-morning I had merely the general aim of doing something fun outside. If I am then offered the chance to go water skiing, I will accept. Moreover, if I am treated as though I have the specific aim of going water skiing, I will not protest or try to spur any course correction. Now, generalizing from this example, if a person has a general aim that G, the person is attributed an aim that F (where F clearly satisfies G), and the person has no other aims that are inconsistent with F, then the person will acquiesce to the attribution of aiming at F and will not protest being treated as aiming at F. Furthermore, if she has an offer of F, which she can accept or decline, she will accept. Crucially, with over-specific aim attributions, such acceptance would not signify full accuracy.

We tend to accept and adopt over-specific aim attributions out of practicality. In fact, adopting an aim that is more specific than what we had in mind is part of our most basic process for satisfying our goals (Richardson 1997). A general aim may be *too general* to be satisfied *at that level of generality*. For example, if I want a writing utensil to scribble down a number, then a pencil, a pen, a crayon, a chalk stick, or a marker would be fine. But the object that actually satisfies my aim will be *one* of these specific types, not somehow a non-committal disjunction of them. In other words, I cannot have a general writing utensil per se, rather only a writing utensil of some more specific type. Hence, if I were offered a pencil I would accept. Thus, in ordinary cases, people accept the attribution of over-specific aims simply because acquisition or realization of the objects of the aims attributed signifies practical success, not failure.[3]

Now we can highlight the crucial epistemic consequence of the tendency to acquiesce to the attribution of an over-specific aim. Consider an ongoing process in which each episode culminates with the attribution of an aim to a person and with the attributer recording the person's response. Suppose that records of those responses will be evidence guiding future aim attributions to this person or others. Since an over-specific attribution will not be protested and indeed may be embraced, it follows that there will be no trace of the inaccuracy constituted by the over-specificity. So, without some deliberate ingenuity to catch the subtle difference, this inaccuracy will register as accuracy. Thus, we can say that over-specific aim attributions *prompt no error signal*.

The second reason inaccuracy due to over-specificity is important regards its practical consequences. Not only are over-specific attributions accepted without prompting an error signal; once accepted, they go on to influence behavior. One reason is simply that aims, once adopted, tend to be stable, and this is a core feature of the operation of practical rationality. As Michael Bratman says of prior intentions and plans: "their retention and non-reconsideration should be treated as the default, but a default that is overridable by certain special kinds of contingencies" (1989, p. 449). Bratman (1992, 2010) argues that such stability is rational for at least three reasons. First, when we have intentions, we tend to act on those intentions, which changes the world in ways that support the continuation of those very intentions. Second, due to limitations of time and mental resources, reconsideration of our intentions is costly. Third, and closely related, habits that limit our tendencies to reconsider our

[3] Although commonsense, this claim is difficult to verify empirically. That is because, given that the claim is true, it entails that what offers an experimental subject would accept would not reveal the specificity of her aims. Introspection about how we make choices may well be our best source of evidence for this claim.

intentions fit well within a larger system of planning agency that increases our effectiveness.

Another reason acceptance of over-specific aim attributions is likely to influence behavior is that specific aims are especially stable. The psychology of implementation intentions shows that specific aims, especially an aim at a specific means to an already established end, tend to be stable (Gollwitzer 1999; Gollwitzer and Schaal 1998).[4] Moreover, studies in goal-setting and motivation have shown that the specificity of a goal increases commitment to it (Wright and Kacmar 1994; Klein et al. 1999).

We are now in a position to see that over-specific aim attributions may constitute self-fulfilling prophecies. If it is indeed true both that people tend to accept over-specific aim attributions and that these aims tend to be stable, then attributing an over-specific aim will tend to bring it about that the person has that very aim. Thus, over-specific attribution of aims is a powerful means of subtly influencing a person. It is a subtle sort of influence because it is gentle—aligning with and guiding, rather than redirecting, a person's aims.[5] Due to this gentleness, it is unlikely to be opposed or even noticed. So, not only is there no error signal at the moment of attribution, the ensuing change is also unlikely to arouse protest.

Presumptuousness and the evidence for aim attributions

To assess the accuracy of an aim attribution is not yet to say anything at all about the evidential basis for the attribution. Our second main distinction regarding the attribution of aims is about evidence and justification for the attributions. With what sort of evidence might a system (whether human or artificial) attribute an aim to a person?

Three categories of justification will be relevant to our purposes here. The first possibility, of course, is that an attribution has no justification or little justification at all. Perhaps it is simply a guess, or, more likely, the aim was imputed purely according to the interests of the attributer independently of whether the person in question actually had that aim. Obviously, this way of attributing aims is not reliably accurate. Furthermore, *some* evidence of the very broad statistical sort—like the general observation that people tend to prefer outdoor recreation on sunny days more than on rainy days—is usually easy to come by. So, since there is usually value in increased reliability, aim attributions with *no* justification are bound to be rare.

Among the attributions that have substantial evidential support, we can identify two rough categories. On one hand, we have those attributions for which the justification is primarily *statistical*. These are based on generalizations about a population applied to the particular person in question. On the other hand, we have attributions that are *individually sensitive*. These attributions are based on facts about the particular individual in question and what aims it makes sense for her to have. Granted, there is no sharp boundary between these two categories. Furthermore, since many attributions are made on the basis of several pieces of evidence, there is no single continuum between the statistically based and the individually sensitive attributions. Nonetheless, we can recognize that some attributions clearly fall more neatly in one category than in the other. For example, think back to the scenarios about Celina's job placement. The attribution by the program administrator was based on general assumptions linking gender to interests and occupational aims. In contrast, Celina's parents drew their conclusion based largely on their knowledge of Celina and her individual tendencies, not through generalizations about a class of superficially similar people. Crucially, this is not to imply that Celina's parents' evidence was completely dispositive, or even that it was probabilistically stronger than the administrator's evidence. Rather, the point is to draw a distinction regarding the *content* of the evidence in relation to the subject of the aim attribution.

To adopt some terminology with more pronounced normative connotations, we can say this: To the extent that an aim attribution was justified statistically, in that its basis was primarily an interpersonal generalization, as opposed to evidence particularly about the person as an individual, the attribution was *presumptuous*. It is presumptuous in that it depends crucially on the presupposition that the person in question will have aims like those of people to whom she bears some other similarity.[6] In invoking the concept of presumptuousness with its negative connotations, I am definitely *not* diminishing the epistemological credentials of the crucial presupposition or the inferences drawn from it. After all, the presupposition is an instance of the assumption on which all inductive inference depends, and interpersonal inductive inferences about persons' aims are indeed credible.

[4] This sort of stability prevents one specific aim from being replaced by an aim at an alternative means of achieving the same general end. However, such specific aims do not necessarily persist independent of the existence and strength of the general aim from which they developed (Gollwitzer and Schaal 1998; Sheeran et al. 2005).

[5] Should we consider this as an infringement of the person's autonomy? It is difficult to say a little bit about autonomy without saying a lot. For now, two observations must suffice. First, because it is influence that aligns with a person's extant aims, a self-fulfilling over-specific aim attribution usurps only marginally a person's authority over what aims to have. Second, because a self-fulfilling over-specific aim attribution may sharpen an aim prematurely (or at least earlier than the person otherwise would have), it reduces the person's ability to plan her life at her own pace.

[6] Presumptuousness has been characterized this way by King (2019), in the context of ethical issues in the development of computerized image recognition systems.

However, ethical status is not the same as epistemological status; the ethics of social cognition is not reducible to the accuracy of social cognition. Several of the ethical issues about presumptuous aim attributions will emerge shortly. For now, the goal is to further clarify the distinction between those attributions that are presumptuous and those that are not.

Consider an aim attribution: *Subject S aims at A*. This attribution is presumptuous if and only if the primary evidence for the attribution consists of these two pieces of information: (1) *S has characteristic C*. (2) *Most people with C aim at A*. For a more concrete example, consider Sara. Sara is an American woman. She is between 25 and 30 years old. Her annual income is in the third quartile of American women in her age bracket. In the last 2 days, she has purchased a swimsuit, a towel, sunglasses, and a chlorine-removing shampoo. Suppose we have data about several thousand people who, like Sara, have the characteristics just listed. And suppose that a vast majority of those people aim to visit an outdoor swimming pool in the next 2 weeks. If, on the basis of all and only this evidence, we attribute to Sara the aim of visiting an outdoor swimming pool in the next 2 weeks, then our aim attribution is presumptuous. That is because nothing about the *content* of Sara's features played a role in the inference, but her similarity to other people did. The inference has the same logical form as this one: *Sara has characteristics #1–#8. A vast majority of people with characteristics #1–#8 have aim A. Therefore, we can conclude that Sara has aim A.* As such, the inference is driven by nothing other than the premise that a particular person will fit the predominant profile of characteristics and aims.

If that commonplace and straightforward inference about Sara was presumptuous, one might wonder, what is the alternative? What would *not* be a presumptuous aim attribution? Well, the very same aim could be attributed to Sara by thinking differently about some of the same information about her. If the information were used, not just to find a pattern she fits, but also to make sense of her, then the ensuing aim attribution would not be so presumptuous. We might reason as follows: First, we know that she bought a swimsuit. Then we can draw on our background knowledge that swimsuits are designed for swimming. So, maybe Sara is planning to go swimming. Also, towels are useful for drying oneself after swimming. Since she has bought sunglasses, which are effective for protecting eyes from the sun, we might also think she is planning to be outside. The aim of going to an outdoor swimming pool, the beach, or some other body of water would explain these bits of information about Sara. But only the aim of going to a chlorinated swimming pool would also explain why she chose that particular shampoo. Hence, we attribute to her the aim of going to an outdoor swimming pool.

This non-presumptuous reasoning was guided by the goal of making sense of Sara, by unifying what we know about her into a coherent portrait of her as an agent. It is true that the reasoning relied on some of the same associations—e.g., between swimsuits and swimming—that might figure in presumptuous reasoning. However, this instance of non-presumptuous reasoning considers these associations in terms of Sara's possible means and ends, instead of as content-neutral, interpersonal, statistical regularities. In short, this inference was *not* presumptuous precisely because it did *not* rely on a premise about what most people resembling Sara aim to do.[7]

At this point, some may object: Wait, how do you know that the tap water in her area is not heavily chlorinated? How do you know that she is not planning to wear that swimsuit in her bathtub? How do you know that the sunglasses are not for driving? Or maybe the sunglasses are a gift? My response to these questions is, of course, that I do not know. These questions suggest alternative hypotheses about Sara's aims, and nothing I have said rules them out. Furthermore, if we were wagering money on what Sara will do, it may very well be that the presumptuous basis for attributing aims would give us more confidence than does the non-presumptuous reasoning just described. To call an aim attribution presumptuous or non-presumptuous is not to assess the strength of the evidence for it, but rather to characterize the assumptions invoked.

There may be a variety of bases for attributing aims *non*-presumptuously. The example of Sara points us to one particularly notable mode of non-presumptuous aim attribution, related to the interpretationist (or interpretivist) theories of mind developed by Donald Davidson and Daniel Dennett (Davidson 1985, 2001; Dennett 1987). According to these theories, correct attributions of aims, beliefs, and other mental states to a person depend on application of an explanatory schema—what Dennett calls *the intentional stance*—which involves attributing mental states to a person as a way of making sense of her history of speech and action and predicting her behavior.[8] Here is how Dennett describes the intentional stance: "[O]ne treats the system whose behavior is to be predicted as a rational agent; one attributes to the

[7] It is one thing to describe non-presumptuous reasoning on which an aim attribution might be based. It would be quite another to automate it. It may turn out that, for some applications, non-presumptuous reasoning is difficult to implement algorithmically. For now, what matters is just to see how the two modes of reasoning differ.

[8] These theories actually make the stronger claim that the mental states *themselves* are, in some sense, interpretation-dependent. I.e., what it is for a person to have these mental states is for her to be such that these are the mental states posited by the most predictive and charitable interpretation of her from the intentional stance (Dennett 1987, 1991; Davidson 2001). For purposes of this article, this strong claim is not required.

system the beliefs and desires it ought to have, given its place in the world and its purpose, and then predicts that it will act to further its goals in the light of its beliefs" (Dennett 1988, p. 496). Attributions of aims according to the intentional stance are not presumptuous. That is because such attributions are not simply projections of patterns observed in other persons.[9]

The intentional stance, which is our ordinary orientation to people, is different from our orientation to other objects in the natural world, and sharply contrasts with what Dennett calls the physical stance: "[O]ne predicts the behavior of a physical system (of stars, tides, volcanoes, DNA molecules) by exploiting information about the physical constitution of the system, and the laws of physics" (Dennett 1988, p. 496). The physical stance is the approach of the natural scientist trying to model and predict the natural world. In contrast to both the intentional stance and the physical stance, presumptuous attributions of aims depend on what we might coin the *statistical stance*. Along the lines of Dennett's descriptions just quoted, we might describe the statistical stance: One treats the system whose behavior is to be predicted as a collection of features, and predicts the values of unknown features by generalizing from similar patterns of features in other systems. The statistical stance can, in principle, be applied to any sort of system, which is no surprise, in light of the broad applicability of statistics and data analytics. When the "systems" in question are persons and the features to be predicted are the mental states of those persons, then the statistical stance can be employed to make the sort of predictions we would otherwise get from the intentional stance.[10] In these cases, since mental states are attributed according to interpersonal generalizations, the attributions are presumptuous. In the above example of Sara, the aim of going to an outdoor swimming pool was attributed, first presumptuously, according to the statistical stance, and then non-presumptuously, according to the intentional stance.

In the next two sections, we will see why it is sometimes morally problematic to employ the statistical stance for social cognition. For now, though, we would be remiss not to recognize the many occasions when it is valuable to apply the statistical stance to humans. Many areas of medical research, particularly clinical trials, exemplify fruitful and unproblematic application of the statistical stance to humans. Such research depends on the premise that individuals who have certain physiological similarities will have other physiological similarities, and the ensuing inferences become the basis for programs of diagnosis, prognosis, and treatment. Although researchers must always negotiate the risks of over-generalization, there is not a problem with generalization per se. The problems with presumptuousness, which *are* problems with generalization per se, arise precisely when the statistical stance is applied to individuals to infer the contents of their aims or other thoughts.

Automated aim attribution and inducement of conformity

Earlier, we examined what it means for an aim attribution to be inaccurate due to over-specificity. Now we have just considered what types of evidence might be the basis for aim attributions, to distinguish between presumptuous and non-presumptuous attributions. In these discussions, we focused on ordinary, everyday cases of interpersonal attributions of aims. But social cognition is not just the purview of humans anymore. Any system capable of attributing mental states and predicting behavior to other systems engages in a variety of social cognition as well. In this section, we will extend the preceding discussions of aim attribution to automated systems that attribute aims, culminating with the suggestion that systems that combine over-specificity and presumptuousness give rise to conformity.

The automated systems in focus are modern recommender systems based on collaborative filtering. Recommender systems are simply automated systems which provide users with recommendations of products, services, or other items (Melville and Sindhwani 2010). Of course, there are many possible goals with which one might design a recommender system, and some of these possible goals may not align with the interests of the user who receives the recommendations. For instance, a system may be designed to maximize profit by recommending the products with the highest profit margin. But, for purposes here, we will focus on the *recommendation problem* as ordinarily formulated, which is the computational challenge of predicting how a user will rate (either by explicit assessment or by choice behavior) items that she has not yet rated (Adomavicius and Tuzhilin 2005). In other words, we will assume the recommender system is

[9] Of course, as both Dennett (1987) and Davidson (1985) recognize, the process of interpretation has to begin somewhere. And if there is nothing else to go on, we may have to begin by positing some basic beliefs and aims. These will be beliefs about truths and aims to bring about what is valuable (Davidson 2004). In exactly the same way, presumptuous attributions of aims may be required when we have nothing else to go on but still must, for whatever reason, attribute an aim to a person.

[10] It is worth emphasizing that, just being geared toward the attribution of aims, desires, beliefs, etc., does not make some method or theoretical orientation count as the *intentional stance*. Differences among the physical stance, the intentional stance, the design stance (Dennett 1987), and the statistical stance do not pertain to the deliverances of each respective method. Rather the differences are in *how* the conclusions were reached and on what basis. Hence, the statistical stance does not become a special case of the intentional stance just in virtue of being applied to the attribution of aims.

intended to provide the user with recommendations which match, as closely as possible, her actual aims.

Recommender systems come in two main varieties. *Content-based* recommender systems make recommendations based on a particular person's *own* past behavior and the extent to which new items are similar to those for which the person has already expressed a preference. In contrast, *collaborative filtering* systems make recommendations according to how *other people* have assessed the items available for recommendation. In practice, these approaches may be combined (Adomavicius and Tuzhilin 2005; Melville and Sindhwani 2010). However, at an abstract level, this distinction between types of recommender systems reflects the distinction in the preceding section between individually sensitive and interpersonal grounds for attributing aims. While content-based systems do not even require multiple users, collaborative filtering systems are inherently presumptuous in the way they depend on inter-user similarity.[11]

To see exactly how collaborative filtering is presumptuous, we must take a closer look.[12] As just noted, recommendations from collaborative filtering systems depend on records of the preferences of other users. In a basic collaborative filtering system, the data consist of item ratings for many combinations of users and items, i.e., ordered triples of *user*, *item*, and *rating*. The relevant items might be anything on which a person might take a recommendation, e.g., songs, movies, books, news stories, social media posts, groceries, restaurants, travel routes, college courses, occupations, romantic partners, etc. The task of such a system is to predict users' ratings of items they have not yet rated.

Consider a user and a new item she has not yet rated. With collaborative filtering, the first step in predicting this user's rating is to identify the users to whom she is most similar.[13] Similarity of two users is based primarily on the similarity of the ratings they have given to the items they have both rated. The basis for similarity may be extended beyond just prior ratings to also include demographic features or personal details (Pazzani 1999). Once the class of similar users has been identified,[14] the next step is to determine how that group of similar users rates the new item, and, on that basis, predict the target user's rating of the item. Ekstrand et al. (2011) provides a helpful high-level description of these systems:

> The fundamental assumption behind this method is that other users' opinions can be selected and aggregated in such a way as to provide a reasonable prediction of the active user's preference. Intuitively, they assume that, if users agree about the quality or relevance of some items, then they will likely agree about other items—if a group of users likes the same things as Mary, then Mary is likely to like the things they like which she hasn't yet seen. (Ekstrand et al. 2011, p. 88)

Thus, predictions of a person's preferences, and, in particular, any attributions of aims that are generated from a collaborative filtering recommender system are guaranteed to be presumptuous. They are not directed toward interpretation of the individual *qua* individual. Rather, they depend essentially on the presupposition that a person will have the same aims as those people to whom she is otherwise similar. They assume the individual in question will be yet another instance of a pattern of features and aims that has emerged in the larger population.

Now we are in a position to observe how such systems may induce and extend the sort of conformity that they already presuppose. The idea is not just that they perpetuate existing conformity, but furthermore that they resolve generality or indeterminacy in a person's aims in the direction of greater conformity. Suppose that a collaborative filtering recommender system is the back-end for an artificial virtual assistant that attributes aims to a person, and accordingly presents her with default courses of action, which she is free to reject or ignore, throughout her day.[15] The level of automation and integration envisioned here is easily foreseeable and only slightly beyond what is presently available. The system might offer intelligent default options for

[11] Content-based recommender systems are not inherently presumptuous, since they do not necessarily even require more than a single user. However, if a content-based approach were adapted to make use of some measure of inter-user similarity, it could potentially generate presumptuous aim attributions as well.

[12] The following descriptions of collaborative filtering systems draw from Adomavicius and Tuzhilin (2005), Schafer et al. (2007), Melville and Sindhwani (2010), and Ekstrand et al. (2011), all of whom agree on all of the basic details relevant here.

[13] This is part of *user-user* (Ekstrand et al. 2011)—also known as *user-based* (Schafer et al. 2007) or *neighborhood-based* (Melville and Sindhwani 2010)—collaborative filtering, where item ratings are predicted on the basis of ratings by similar users. An alternative is *item-item* (Ekstrand et al. 2011)—also known as *item-based* (Schafer et al. 2007; Melville and Sindhwani 2010)—collaborative filtering, according to which earlier ratings are used to group items (rather than users) by similarity. With item-item systems, items are grouped according to whether they are highly rated by the same group of users. Although we will focus on user-user collaborative filtering, it is worth noting that both user-user and item-item methods yield presumptuous recommendations, since both use interpersonal similarity in one area to predict interpersonal similarity in another.

[14] There need not be a sharply circumscribed group of similar users. Instead every user can be taken into account, but with their ratings weighted according to how similar they are to the target user.

[15] Presenting a user with default courses of action which she is free to ignore or reject would count as a *nudge* (Thaler and Sunstein 2008). Hence, the worries such a system raises regarding user autonomy are less severe than the worries we might have with a more coercive system. This gives us room to focus on the particular moral concerns raised by presumptuousness and over-specificity, while bracketing further issues about autonomy.

food, comfort, news, entertainment, and travel—i.e., default meals, default ambiance, default articles and stories, default music and videos, default routes and destinations. These defaults could be offered as suggestions, perhaps via push notifications, or the items themselves could be actually provided by the likes of smart home automation products, driverless cars, etc.

The technological set-up just laid out embodies the two key characteristics of aim attributions described in previous sections. First, any aim attributions generated by such a system will be presumptuous, based as they are on the assumption that the target user has aims like other users who are like her. Second, the aims attributed will usually be over-specific. That is simply because the real options (like particular products and services) that are actually available to be offered are necessarily particular and concrete, not general or abstract. Consider: To enable travel, a navigation system or autonomous vehicle cannot simply opt for country roads rather than the highway; it must select a particular country road. Similarly, music of a particular genre cannot be provided as a default without also providing particular tracks or pieces that exemplify that genre. Thus, a high degree of specificity is not a contingent fact about these systems, but rather an inherent characteristic of the provision of actionable recommendations or fully concrete default options.[16] Insofar as the user's antecedent aims are not so specific, the system's recommendations will be over-specific.

In light of this presumptuousness and over-specificity, a mechanism for the inducement of conformity is apparent. When the system attributes an over-specific aim to the user (by offering her a specific item that satisfies one of her general aims), the user will likely acquiesce and is unlikely to protest. So this aim attribution, insofar as it is over-specific but not simply inaccurate, will prompt no error signal; there will be no signal to stop, adjust, or otherwise correct the system. And since the item offered satisfies the user's antecedent general aim, the new present intention to proceed with the item is likely to stabilize. Thus, the aim attributions are self-fulfilling: The person's general aims evolve to more specific aims according to the specific attributions by the system she is using.

Since the system in question attributes aims according to collaborative filtering, it attributes aims presumptuously. So, since the system is guided by the premise that this user will have aims like other users, this user's aims will sharpen to conform with the specific aims of others. Hence, when the presumptuous attribution of an aim triggers its own fulfillment, a prediction of conformity becomes a promotion of conformity.

I have deliberately constructed the set-up just described to allow a clear interface between modern recommender systems and the foregoing premises about over-specific and presumptuous attributions of aims. However, mechanisms for the inducement of conformity will not always be so straightforward. Hence, the technological scenario described above is offered as something like a proof of concept. I conjecture that in more complicated scenarios, the conformity effects will be less direct and less pronounced for the individual users involved, but similarly significant in aggregate. If that is correct, then inducements to conformity driven by automated presumptuousness will be subtle, and, for that reason, potentially more likely to pass unnoticed.

This section described a straightforward mechanism through which collaborative filtering recommender systems may contribute to interpersonal conformity in our aims. This is one salient case among wider varieties of *deindividualization* due to the application of large-scale analytics to personal data (Vedder 1999), and we should find it concerning. Many of us, I expect, are inclined to follow Mill in holding that employing our individual faculties to choose our own individual paths is a deep, perhaps foundational, source of value in our lives (Mill 1859/2003). If so, then we should avoid mechanisms that promote conformity and discourage their use. However, there is room for disagreement about the goodness or badness of conformity, and I have no pretense of settling that issue. For present purposes, it will suffice to note that conformity, and its antithesis, individuality, are morally relevant categories. Therefore, the tendency to induce conformity should be investigated, considered, and weighed, when designing and deploying such systems.

The broader moral significance of automated presumptuousness

We have seen that automating particular sorts of inferences through recommender systems may induce interpersonal conformity. This is due, in part, to the fact that these systems operate on the basis of the presumptuous assumption that individuals will have aims similar to those of people who are otherwise similar. Putting aside the question of whether these systems do indeed perpetuate conformity, there is a more basic worry about how these systems operate. If presumptuous aim attributions are intrinsically morally problematic, even if just moderately so, then acceleration and proliferation of such attributions by automated systems is morally objectionable, independently of the eventual social consequences. In this section, I will suggest that presumptuous aim attribution is indeed inherently morally problematic.

[16] Over-specificity can be mitigated by recommending a list of options instead of a single one. Although each option would be over-specific, the full list may better reflect a user's general aim. Of course, whether this is possible depends on whether the use case requires automatic selection of a single option (as in the case of autonomous navigation).

Table 1 Morally significant categories of aim attribution

	Simply inaccurate	Over-specific	Accurate
Presumptuous attribution	(#1) Mistaken deindividualization	(#3) Inducement of conformity	(#5) Statistical fidelity
Individually sensitive attribution	(#2) Misapprehended individuality	(#4) Cooperative individualism	(#6) Interpersonal understanding

We will consider how presumptuousness of an aim attribution combines with its accuracy or inaccuracy to affect how we evaluate it. Recall our four-way distinction regarding accuracy: Aim attributions may be accurate, inaccurate because over-specific, inaccurate because under-specific, or simply inaccurate. To avoid adding complications, we will continue to ignore under-specificity. Next, regarding evidence, we had a three-way distinction among attributions made without any evidence, attributions made presumptuously, and individually sensitive attributions. Since they are unusual, we will continue to ignore attributions without any evidence. That leaves us with a three-way distinction and a cross-cutting two-way distinction, and so six cases for evaluation, as shown in Table 1.

I have deliberately given each of the six cases a provocative label. The only one of these labels I can hope to thoroughly justify is "inducement of conformity" for box #3, on the basis of the argument of the preceding section. The rest are intended primarily to be suggestive, to direct our attention to broader worries about presumptuous attributions of aims. As we examine each column in turn, we will see that our evaluations are sensitive not just to the accuracy of the aim attributions but also to the evidence on which they are based.

In box #1, we have the application of an interpersonal generalization to yield a simply inaccurate conclusion about a person. In contrast, in box #2, the simple inaccuracy is due to the fact that we sometimes have incorrect information about each other. To see the distinction and what difference it makes, consider an alternate version of the Celina scenarios. Suppose that during her travels Celina faced some unexpected challenges but rose to the occasion. Further suppose that, as a result, she quickly developed a self-reliance, confidence, and assertiveness that neither she nor others had recognized in her before. Celina immediately values these traits and embraces them as core elements of her character. Then, before returning home, she seeks opportunities to exercise these traits, and the parking monitor job would have been especially attractive to her. Nevertheless, continue to suppose that both Celina's parents and the program administrator, oblivious to recent events in Celina's life, attribute to her the aim of selling university merchandise according to the rationales described earlier.

In this alternate version of the story, the aim attributed by both Celina's parents and the program administrator is *simply inaccurate*. In this version, as in the original, both attributions are made (due to the situation) without actually consulting her; hence, the attributions are equally paternalistic (if paternalistic at all). Now we can notice that, despite these similarities, Celina's parents' inference seems sensible, while the administrator's inference seems objectionable. Both inferences were inductive, reasoning from a pattern in seen instances, to a conclusion about a new case. However, the administrator's basis was an interpersonal pattern of aims and interests, while Celina's parents relied on Celina's own individual history of aims and interests. Although Celina's parents lacked the most up-to-date information about what sort of individual Celina had become, they still treated her as an individual. In contrast, the administrator's reasoning treats Celina primarily as a member of a group, rather than as an individual with a particular set of attributes. Thus, his inference exemplifies a characteristic that makes some stereotypes objectionable (Blum 2004),[17] and constitutes a sort of deindividualization (Vedder 1999). Hence box #1 appears objectionable in a way that box #2 does not.

We move now to boxes #3 and #4. The discussion of the conformity-inducing properties of collaborative filtering recommender systems belongs in box #3. We have seen that these cases are worrisome because these systems realize a mechanism for perpetuating a presupposition of conformity by directing the development of our general, not-yet-refined aims. In contrast, notice that the aim attributions of box #4 are less worrisome. In fact, the box #4 attributions of over-specific aims may be quite desirable. Due to shortcomings of our information, experience, or imagination, we are not always well-equipped to further specify our own aims all by ourselves. The advice of those who know us well—those who are sensitive to the traits that make us unique individuals—is among our most valuable resources as we grow and evolve as persons.[18] Recognizing the value

[17] Though it raises some of the same moral concerns as stereotyping, presumptuous attribution of mental states is broader than stereotyping in one way and narrower in another. It is broader in that the generalizations involved need not invoke well-recognized social categories. It is narrower in that it applies only to the attribution of mental states, not other characteristics (at least not directly).

[18] The advice of those who know us well carries dangers of its own. Among the risks here is the premature specification of our aims, which may short-circuit the trial-and-error process by which we might otherwise develop as individual agents and persons.

of such assistance is to mark a limit to the individualism that grounded our complaints about the presumptuous attributions of over-specific aims in box #3. A moderate, cooperative individualism that objects to sly inducements of conformity should not be conflated with a fend-for-yourself ethos that rejects the benefits of close relationships formed in community with others. Hence, box #4 is largely free of the worries aroused by box #3.

Finally, turn to boxes #5 and #6. Accurately attributing to a person an aim she actually has, based on a rich knowledge of her as an individual, can be wonderful. It is a nourishment to and a fruit of the deepest relationships we have. These aim attributions, which we find in box #6, manifest a kind of interpersonal understanding that is regrettably rare when our connections are too superficial.

Compare these to the accurate attributions made presumptuously in box #5. Here the specificity of the aims attributed matches the specificity of the person's actual aims. In the case of product recommendation, the likely reason an attribution would be accurate instead of over-specific would be that the person actually wanted something very specific, and the interpersonal statistical inference yielded a recommendation of that very thing. These high-fidelity, presumptuous attributions may strike us as almost magical, or perhaps even creepy, precisely because the attributer did not have the right information to make sense of such an accurate attribution. The magic ingredient, of course, would be the presupposition of interpersonal similarity that drives the statistical inference. This was the element that we regarded as objectionable in box #1. We are likely to find box #5 less objectionable than box #1, though, because of the value of accuracy, as well as the delight we take in the impression that we have been understood, even if (as in box #5) the understanding was not of us as individuals but primarily of the patterns we instantiate. Only in box #6, where the attribution is both accurate and non-presumptuous, is there genuine interpersonal understanding.

In the preceding section, we worried that automated recommendation systems may have the effect of diminishing individuality across a population. In working through the array of cases in Table 1, we have seen how presumptuous attributions of aims efface individuality in a more direct way. People and systems that operate presumptuously fail to *treat* persons as individuals. They focus on trends of a larger group and consider the individual person only derivatively. This is the moral problem of presumptuousness, which we should hope not to be accelerated and proliferated through automation.

To deal with a person by treating her as a member of a group is not to deal with her as an individual person per se. Rather, it is to deal with her more like an object among other objects. Along these lines, we close with a connection to Peter Strawson's distinction between responding to a person with *objective attitudes* and responding to the person with *reactive attitudes*.

> To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided… The objective attitude… cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships. (Strawson 1974, p. 9)

Even if good lives do not require constant active involvement in full-fledged interpersonal relationships, we may nevertheless worry that our lives are impoverished when we are *too often* subject to objective attitudes. As artificial systems become increasingly involved in our lives and assume roles that other humans have, our relationships built around reactive attitudes may be displaced by relationships dominated by the statistical stance and presumptuous aim attributions. If so, then we should worry that, in being increasingly treated as mere objects, we increasingly come to be mere objects.

Acknowledgments I wish to thank the following people for feedback on earlier drafts of this article: Philip Brey, Catherine Greene, Hinda Haned, Michael Kühler, Ana Lucic, Alan Rubel, Isabel Taylor, Bart Voorn, Indy Wijngaards, and especially Mayli Mertens. I am grateful to the following people for discussion of the topic: Andréa Atkins, Justin D'Arms, Brandt van der Gaast, Tommy Hofkamp, Michael Milona, Shyam Nair, Sander Voerman, Maria Lisa de Vries, and Clark Woods. The article also benefited from discussions with members of the Tech and Values group at the University of Twente and attendees of the 2017 OZSW Annual Conference.

Funding Work on this article was supported by the Netherlands Organisation for Scientific Research (NWO), as part of the New Science of Existential Well-Being (NEWEL) project (NWO project number 652.001.003).

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 6,* 734–749.

Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.

Blum, L. (2004). Stereotypes and stereotyping: A moral analysis. *Philosophical Papers, 33*(3), 251–289.

Bratman, M. (1989). Intention and personal policies. *Philosophical Perspectives, 3,* 443–469.

Bratman, M. (1992). Planning and the stability of intention. *Minds and Machines, 2*(1), 1–16.

Bratman, M. (2010). Agency, time, and sociality. *Proceedings and Addresses of the American Philosophical Association, 84*(2), 7–26.

Davidson, D. (Ed.). (1985). Belief and the basis of meaning. In *Inquiries into truth and interpretation* (pp. 141–154). Oxford: Oxford University Press.

Davidson, D. (Ed.). (2001). Mental events. In *Essays on actions and events* (2nd ed., pp. 207–225). New York: Oxford University Press.

Davidson, D. (Ed.). (2004). Expressing evaluations. In *Problems of rationality* (pp. 19–37). Oxford: Oxford University Press.

Deci, E. L. (1992). On the nature and functions of motivation theories. *Psychological Science, 3*(3), 167–171.

Dennett, D. (Ed.). (1987). True believers. In *The intentional stance* (pp. 13–35). Cambridge, MA: MIT Press.

Dennett, D. (1988). Precís of the intentional stance. *Behavioral and Brain Sciences, 11*(3), 495–505.

Dennett, D. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27–51.

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction, 4*(2), 81–173.

Fiske, S. T., & Taylor, S. E. (2016). *Social cognition: From brains to culture* (3rd ed.). Thousand Oaks, CA: Sage Publishing.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493–503.

Gollwitzer, P. M., & Schaal, B. (1998). Metacognition in action: The importance of implementation intentions. *Personality and Social Psychology Review, 2*(2), 124–136.

Hamilton, D. L., Stroessner, S. J., & Driscoll, D. M. (1994). Social cognition and the study of stereotyping. In P. G. Devine, D. L. Hamilton, & T. M. Ostrom (Eds.), *Social cognition: Impact on social psychology* (pp. 291–321). San Diego, CA: Academic Press.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995.

King, O. C. (2019). Machine learning and irresponsible inference: Morally assessing the training data for image recognition systems. In D. Berkich & M. V. d'Alfonso (Eds.), *On the cognitive, ethical, and scientific dimensions of artificial intelligence* (pp. 265–282). Heidelberg: Springer.

Klein, H. J., Wesson, M. J., Hollenbeck, J. R., & Alge, B. J. (1999). Goal commitment and the goal-setting process: Conceptual clarification and empirical synthesis. *Journal of Applied Psychology, 84*(6), 885–896.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Locke, E. A., & Latham, G. P. (2013). Goal setting theory, 1990. In E. A. Locke & G. P. Latham (Eds.), *New developments in goal setting and task performance* (pp. 3–15). London: Routledge.

Melville, P., & Sindhwani, V. (2010). Recommender systems. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 829–838). Heidelberg: Springer.

Mill, J. S. (1859/2003). *Utilitarianism and on liberty* (2nd ed.) Mary Warnock (Ed.). Oxford: Blackwell Publishing.

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review, 13*(5–6), 393–408.

Richardson, H. S. (1997). *Practical reasoning about final ends*. Cambridge: Cambridge University Press.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web* (pp. 291–324). Heidelberg: Springer.

Schroeder, T. (2004). *Three faces of desire*. Oxford: Oxford University Press.

Sheeran, P., Webb, T. L., & Gollwitzer, P. M. (2005). The interplay between goal intentions and implementation intentions. *Personality and Social Psychology Bulletin, 31*(1), 87–98.

Smith, M. (1994). *The moral problem*. Oxford: Blackwell Publishing.

Strawson, P. F. (Ed.). (1974). Freedom and resentment. In *Freedom and resentment and other essays* (pp. 1–27). London: Methuen.

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Vedder, A. (1999). KDD: The challenge to individualism. *Ethics and Information Technology, 1*(4), 275–281.

Wright, P. M., & Kacmar, K. M. (1994). Goal specificity as a determinant of goal commitment and goal change. *Organizational Behavior and Human Decision Processes, 59*(2), 242–260.