Should you let AI tell you who you are and what you should do?

Muriel Leuenberger

Accepted Manuscript. Forthcoming in *Living with AI: Moral Challenges*. Edited by David Edmonds. Oxford: Oxford University Press.

Your phone and its apps know a lot about you. Who you are talking to and spending time with, where you go, what music, games, and movies you like, how you look, which news articles you read, who you find attractive, what you buy with your credit card and how many steps you take. Personal information about individual preferences, characteristics, and actions has turned digital. Nearly everything you might want to know about a person is available or can be inferred from stored 1s and 0s. This information is already being exploited to sell us products, services, or politicians. Online traces allow companies like Google or Facebook to infer your political opinions, consumer preferences, whether you are a thrill-seeker, a pet lover, or a small employer, how probable it is that you will soon become a parent, or even whether you are likely to suffer from depression or insomnia.

With the use of artificial intelligence and the further digitalization of human lives, it is no longer unthinkable that AI might come to know you better than you know yourself. The personal user profiles AI systems generate could become more accurate in describing their values, interests, character traits, biases, or mental disorders than the user themselves. Already, technology can provide personal information that individuals have not known about themselves. Such elaborate personal profiles might enable AI to tell you what you really want – not just which movies or music fit your taste – but who you want to vote for, which job would suit you, or which potential partner you should date. Yuval Harari[1] exaggerates but makes a similar point when he claims that it will become rational and natural to pick the partners, friends, jobs, parties, and homes suggested by AI. AI will be able to combine the vast personal information about you with information about psychology, relationships, employment, politics, and geography, and it will be better at simulating possible scenarios regarding those choices.

Knowing yourself and improving your life choices is advantageous in many ways. Self-knowledge is instrumentally useful since it saves a lot of time and stress by helping you to make

[1] Y. N. Harari, *Homo Deus: A Brief History of Tomorrow* (Harvill Secker, 2016).

plans and decisions that are more likely to enhance your well-being. It is also good because it makes you a more reliable co-operator. What's more, you will cost the health system less if you know more about your body and mind. Some philosophers even argue that you owe it to yourself to know yourself. According to Kant, for instance, this duty to oneself grounds all moral duties. Self-knowledge mitigates the tendency to deceive yourself into believing that what you want is morally right. Even if AI did not increase self-knowledge but helped you make better personal decisions without fully understanding them, your well-being would likely increase. Society at large would probably also profit, because you would, for instance, vote for a political party that actually represented your interests instead of the one with the bigger advertisement budget.

So it might seem that an AI that lets you know who you are and what you should do would be great, not just in extreme cases, à la Harari, but more prosaically for common recommendation systems and digital profiling.

I want to raise two reasons why it is not.

Trust

How do you know whether you can trust an AI system? How can you be sure whether it really knows you and makes good recommendations for you? Imagine a friend telling you that you should go on a date with his cousin Alex because the two of you would be a perfect match. When deciding whether to meet Alex you reflect on how trustworthy your friend is. You may consider your friend's reliability (is he currently drunk and not thinking clearly?), competence (how well does he know you and Alex, how good is he at making judgements about romantic compatibly?), and intentions (does he want you to be happy, trick you, or ditch his boring cousin for an evening?). To see whether you should follow your friend's advice you might gently interrogate him: why does he think you would like Alex, what does he think you two have in common?

This is complicated enough. But judgements of trust in AI are more complicated still. It is hard to understand what an AI really knows about you and how trustworthy its information is. Many AI systems have turned out to be biased – they have, for instance, reproduced racial and sexist biases from their training data – so we would do well not to trust them blindly. Typically, we can't ask an AI for an explanation of its recommendation, and it is hard to assess its reliability, competence, and the developer's intentions. The algorithms behind the predictions, characterizations, and decisions of AI are usually company property and not accessible by the user. And even if this information were available, it would require a high degree of expertise to comprehend it. How do those purchase records and social media posts translate to character traits and political preferences? Because of the much-discussed opacity, or 'black box' nature of some AI systems, even those proficient in computer science may not be able to understand an AI system fully. The process of how AI generates an output is largely self-directed (meaning it generates its own strategies without following strict rules designed by the developers), and difficult or nearly impossible to interpret.

Moreover, the novelty of AI and the fast development of new versions, updates, and improvements make it hard to rely on methods to establish trust that require longer timeframes. It just takes a few minutes to check whether Spotify recommendations capture your musical taste. But to know whether the system is competent at making recommendations that affect a person over long time spans, such as career or partner choices, we need long-term data. This data won't be available when the technology is implemented; possibly it will never be available for any specific model at the time of use.

The complex, statistical categorizations made by AI further confuse matters. To make recommendations for, or decisions about, individuals, AI commonly creates groups based on opaque statistical correlations, such as "probably not suffering from a sexually transmitted disease" or "likely to be an atheist".[2] But it often remains unclear to the individual how they ended up in a group, and what they have in common with other members of the group. Moreover, the label they are identified with may appear alien and disconnected from their lived experience because it does not have a corresponding group-identity in real life. Should you feel affiliated with the group of novelty-sweater buyers in which the AI has placed you? It becomes difficult to understand what those characterizations mean, to evaluate the appropriate reaction, or to challenge them. Because it is unclear how the AI system generates user profiles, groups, and recommendations, it is hard know what to do with this information. Should you trust it, discard it, or fight it?

As long as it is just a song recommendation, or perhaps Google labelling you as a thrill-seeker, this might not be a pressing issue. But in some cases, the stakes are substantially higher. AI is being used in job recruiting and selection, in medical decision-making, and in the justice system to assess recidivism risk or for predictive policing. An example of the latter is the Chicago Strategic Subject List or Heat List implemented by the Chicago Police Department. Between 2012 and 2019, they used an algorithm to list and score people according to their probability of

[2] K. de Vries 'Identity, profiling algorithms and a world of ambient intelligence', in *Ethics and Information Technology*, 12/1 (2010), 71-85.

being involved in a homicide or non-fatal shooting – either as a victim or a perpetrator.[3] Police then approached high-scoring individuals to inform them that they were deemed likely to be involved in a shooting and that they, the police, were keeping an eye on them. What should those individuals do with this information? Should they accept it as a reasonable representation of who they are and what their future holds?

The opacity of AI profiling and decision-making leaves few opportunities for negotiation. On what grounds can you disagree with an opinion that is presented with the authority of scientific fact, computational science, and statistics, without knowing how this opinion has been generated? How would someone who objected to being on the Chicago Heat List, and to being described as probable future shooters or gun victims go about contesting this claim, not just on legal grounds, but to convince friends, family, and ultimately themselves that this is not who they are? The question who really knows you better, yourself or AI, and who has the authority about who you are, remains hard to answer as long as it is unclear whether, in which situations, and to what degree you can trust the AI. In fact, in the case of the Chicago Heat List, a study exposed how ineffective the program was in combatting crime.

Create yourself!

Even if we had a reasonably trustworthy AI, a second ethical concern would remain. An AI that tells you who you are and what you should do is based on the idea that your identity is something you can discover – information you or an AI may access. Who you really are and what you should do with your life is accessible through statistical analysis, some personal data, and facts about psychology, social institutions, relationships, biology, and economics. But this view misses an important point: we also *choose* who we are. You are not a passive subject to your identity – it is something you actively and dynamically create. You develop, nurture, and shape your identity. This self-creationist facet of identity has been front and centre in existentialist philosophy, as exemplified by Jean-Paul Sartre. Existentialists deny that humans are defined by any predetermined nature or "essence". To exist without essence is always to become other than who you are today. We are continually creating ourselves and should do so freely and independently. Within the bounds of certain facts – where you were born, how tall you are, what you said to your friend yesterday – you are radically free and morally required to construct your own identity and define what is meaningful to you. Crucially, the goal is not to unearth

[3] A major problem with the use of AI in the justice system is that it can reinforce or lead to discrimination. For more on this see chapter … (Binesh Hass) in this volume.

the one and only right way to be but to choose your own, individual identity and take responsibility for it.

This self-creation occurs in two, mutually influencing ways: via action and via interpretation. First of all, you decide who you are through your actions. By acting, you create facts about yourself. You become a person who has done X or refused to do Y. Either you are a person who comes to help his friend in need, or you are not. Moreover, by acting in certain ways and exposing yourself to certain people, emotions, or environments, you gradually shape yourself and change your views, values, beliefs, and/or desires. This might be done in a self-conscious attempt at changing yourself, but we often shape ourselves less purposefully. By choosing a career, for instance, you expose yourself to a certain type of person and certain situations which will influence who you are, whether you like it or not.

At the same time, you also create yourself by conceptualizing and interpreting yourself. The crude facts about people and their actions do not by themselves define their identity. They provide the raw material for self-definition. Figuring out what the facts *mean* is a matter of interpretation, and to an extent, that interpretation is up to you. Thus, defining yourself involves interpreting actions, understanding what motives drove them and how they connect to overarching intentions, evaluating and organizing personal information, finding patterns, labelling yourself, and making choices about what is important to your identity and what is irrelevant, or which events, social groups, actions, or achievements are defining you and in what way. Self-defining actions and crude facts leave room for creative, interpretive self-definition. To some degree, it is up to you whether you find your teenage celebrity crush embarrassing or embrace it as part of who you are, whether your nationality is a contingent fact about you or something you deeply identify with, or whether doing philosophy research is just a job or a calling. This self-image you create is action-guiding. You get to plan and lead a life in accordance with who you take yourself to be. This interpretive dimension of self-creation feeds back into the agentive one. At the same time, creating yourself through interpretation takes your actions, as well as other facts about yourself, as a starting point.

AI can give you an external, quantified perspective which can act as a mirror and suggest courses of action. But you should stay in charge and make sure that you take responsibility for who you are and how you live your life. An AI might state a lot of facts about you, but it is your job to find out what they mean to *you* and how you let them define you. The same holds for actions. Your actions are not just a way of seeking well-being. Through your actions, you choose what kind of person you are. Blindly following AI entails giving up the freedom to create yourself and renouncing your responsibility for who you are. This would amount to a moral failure.

There is great value in the process of choosing for yourself. The deliberation and experimentation involved in making your own choices is an exercise in self-understanding and self-making. Through this process, you learn and decide what your opinions are or why you like X and identify as Y. You ascribe meaning to actions and facets of your identity when you deliberate and choose because they become connected to insights, beliefs, values, and overarching intentions. You may come to realise that you chose a career because you value the security and stability it provides. Through this process, you also understand the meaning of this job for you, as a source of security. If an AI made this choice on your behalf, the job might not carry this kind of meaning (even though you could reconstruct and ascribe this meaning in retrospect). Moreover, by choosing for yourself, you can be sure that your own interests are guiding the process. When you rely on others or on AI, different interests can come into play. Companies and government institutions that use and provide AI profiling and recommendation systems pursue their own concerns, which might be detrimental to your own well-being. While it might be financially beneficial for YouTube if you watch more of their recommended videos, it would be bad for your well-being if you are thereby politically radicalized and start believing in conspiracy theories.

Ultimately, relying on AI to tell you who you are and what you should do can stunt the skills necessary for independent self-creation. If you constantly use an AI to find the music, career, or political candidate you like, you might eventually forget how to do this yourself. AI may deskill you not just on the professional level but also in the intimately personal pursuit of self-creation. You might profit from using AI as a useful tool to make suggestions for your decision-making and self-interpretation. But given the value of self-creation, you should be careful to retain the skills for doing this on your own. Choosing well in life and construing an identity that makes you happy is an achievement. By subcontracting this power to an AI, you gradually lose responsibility for your life and ultimately for who you are. You no longer deserve blame or praise for your choices and tastes (or, in the extreme case, for how your life evolves) because you are only a passive recipient of choices made by a computer.

Of course, we often make bad choices. But this has an upside. By exposing yourself to influences and environments which are not in perfect alignment with who you currently are you develop. Moving to a city that makes you unhappy could disrupt your usual life rhythms and nudge you, say, into seeking a new hobby. This would change you into someone with novel preferences and routes to well-being. Recommendation systems look at who you are now. Constantly relying on AI recommendation systems might calcify your identity. This is, however, not a necessary feature of recommendation systems. In theory, they could be designed to broaden the user's horizon, instead of maximising engagement by showing customers what they already like. In practice, that's not how they function.

This calcifying effect is reinforced when AI profiling becomes a self-fulfilling prophecy. It can slowly turn you into what the AI predicted you to be and perpetuate whatever characteristics the AI picked up. By recommending products and showing ads, news, and other content, you become more likely to consume, think, and act in the way the AI system initially considered suitable for you. The technology can gradually influence you such that you evolve into who it took you to originally be. If Google thinks you like SUVs and bombards you with SUV ads and content, you are more likely to develop a preference for SUVs. Because recommendation systems perpetuate identified characteristics and those characteristics are often modelled after patterns found in other people – what does someone want who is your age, who visits the kind of locations you regularly go to, and who has your type of purchase record and news consumption want – they level out outliers. While those recommendations aim to be individualized and uniquely customized for you, they often build on universals and averages, on data about what others with your characteristics do, want, and chose (this is called collaborative filtering). They do not make everyone the same, but they do make you more likely to resemble others within your specific bubbles.

The Chicago Heat List became a self-fulfilling prophecy for at least one person on the list.[4] Robert McDaniel had no felonies or violent offences on his criminal record. He had no idea why he was deemed likely to be involved in a shooting or how he could change that. McDaniel was repeatedly visited by the police and was put under constant surveillance after the algorithm gave him a high score on the Heat List. As a result, people who noticed the regular police interactions suspected him to be an informant and shot at him on two separate occasions. He was injured but survived both incidents.

*

You may sometimes wish for someone to tell you what to do or who you are. But, as we have seen, this comes at a cost. It is hard to know whether or when to trust AI profiling and recommendation systems. More importantly, by subcontracting decisions to AI, you may fail to meet the moral demand to create yourself and take responsibility for who you are. In the process, you may lose skills for self-creation, calcify your identity, and cede power over your identity to companies and government. Those concerns weigh particularly heavy in cases involving the most substantial decisions and features of your identity. But even in more

[4] M. Stroud, *Heat Listed* (The Verge, 2021) https://www.theverge.com/c/22444020/chicago-pd-predictive-policing-heat-list.

mundane cases, it would be good to put recommendation systems aside from time to time, and to be more active and creative in selecting movies, music, books, or news. This in turn, calls for research, risk, and self-reflection.

Recommendation for further reading

J. Cheney-Lippold, *We Are Data: Algorithms and The Making of Our Digital Selves* (New York University Press, 2017).