Philosophy & Technology (2021) 34 (Suppl 1):S65–S90
https://doi.org/10.1007/s13347-020-00406-7

RESEARCH ARTICLE

Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias

Ying-Tung Lin[1] · Tzu-Wei Hung[2] · Linus Ta-Lun Huang[2,3]

Received: 2 October 2019 / Accepted: 26 May 2020 / Published online: 3 July 2020

Abstract

This paper focuses on the potential of "equitech"—AI technology that improves equity. Recently, interventions have been developed to reduce the harm of implicit bias, the automatic form of stereotype or prejudice that contributes to injustice. However, these interventions—some of which are assisted by AI-related technology—have significant limitations, including unintended negative consequences and general inefficacy. To overcome these limitations, we propose a two-dimensional framework to assess current AI-assisted interventions and explore promising new ones. We begin by using the case of human resource recruitment as a focal point to show that existing approaches have exploited only a subset of the available solution space. We then demonstrate how our framework facilitates the discovery of new approaches. The first dimension of this framework helps us systematically consider the analytic information, intervention implementation, and modes of human-machine interaction made available by advancements in AI-related technology. The second dimension enables the identification and incorporation of insights from recent research on implicit bias intervention. We argue that a design strategy that combines complementary interventions can further enhance the effectiveness of interventions by targeting the various interacting cognitive systems that underlie implicit bias. We end with a discussion of how our cognitive interventions framework can have positive downstream effects for structural problems.

Keywords Implicit bias · Decision support · Augmented decision · Fairness · AI4SG · Artificial intelligence

The authors contribute equally to this paper

✉ Linus Ta-Lun Huang
linushuang@ucsd.edu

Ying-Tung Lin
linyingtung@gmail.com

Tzu-Wei Hung
htw@sinica.edu.tw

Extended author information available on the last page of the article

1 Introduction

Implicit bias refers to a type of automatic stereotype or prejudice that affects our opinions, decisions, and behaviors (Brownstein 2019). Its harmful impacts include discrimination against individuals based on factors such as race, ethnicity, gender, and social class (Dunham and Leupold 2020). However, recent research reveals a growing consensus that implicit bias is subserved by multiple types of interacting cognitive mechanisms (Huebner 2016; Schwitzgebel 2013; Brownstein 2019). The complex nature of implicit bias may partly explain why existing bias-reduction interventions, which target only one or a small number of cognitive mechanisms, have limited effectiveness. This might also explain why, in turn, even the most successful interventions only seem to work in the short term (Lai et al. 2014, 2016; Lai and Banaji 2019; Liao and Huebner 2020). Given the pernicious effect of implicit bias, we urgently need to explore new intervention strategies.

One emerging intervention strategy relies on artificial intelligence (AI), the human-made computational systems capable of solving specific problems. The extensions of AI include machine learning, smart robotics, computer vision, virtual agents, etc. Chamorro-Premuzic (2019), for instance, argues that AI systems can be programmed to ignore information that is irrelevant to certain decisions (e.g., a job applicant's gender in hiring a computer programmer). This allows AI to analyze only information that is relevant to the job requirements (e.g., programming skills) in order to reach an unbiased decision. However, AI systems have been found to perpetuate bias, due to either the unintended consequences of algorithmic design or problematic data (Obermeyer et al. 2019; Richardson et al. 2019; Sweeney 2013). Pessimism about AI-assisted approaches to eliminating implicit bias prevails.

We believe this pessimism is premature and argue that "equitech"—AI technology for improving equity—has untapped potential. In this paper, we provide a framework for exploring innovative AI-assisted interventions that can effectively reduce the harm of implicit bias. "Section 2" begins with a discussion of implicit bias that emphasizes its complex cognitive nature. In "Section 3," we introduce our framework and use human resource recruitment as a case study. Using this framework, we show that existing approaches to reducing bias in hiring processes face various limitations and overlook opportunities offered by advancements in technology and cognitive research. In "Section 4," we demonstrate the utility of our framework for discovering novel approaches. Specifically, we show that our framework helps incorporate recent developments in AI-related technologies, as well as insights from philosophical and empirical research on implicit bias. In "Section 5," we show how our framework helps to design and combine interventions in a complementary way. Two implications follow from this. The first is that, given implicit bias's complex nature, it is best to combine complementary interventions to target the multiple interacting mechanisms that underlie it. Second, after endorsing the view that implicit bias involves the dynamic interplay of cognitive, social, and physical factors (Liao and Huebner 2020; Soon 2019), we suggest, in "Section 6," that interventions targeting multiple cognitive factors can have a strong positive impact on more structural problems. We conclude with cautious optimism that, despite some unresolved limitations, the future of equitech is promising.

2 Complexity of Implicit Bias

Implicit bias is an automatic form of stereotype or prejudice people unintentionally act on (Brownstein 2019).[1] It is contrasted with explicit bias, which one is aware of and/or can intend to act on accordingly.[2] Implicit bias can dispose us toward epistemically flawed beliefs and/or morally wrong decisions and actions (Holroyd and Sweetman 2016). For example, black students are often rated as less academically capable than their identically performing white peers (Hodson et al. 2002). Meanwhile, in the domain of gender, research suggests that most scientists unconsciously associate science with men (Régner et al. 2019). Implicit bias is a prevalent and pernicious phenomenon.

Notably, there is a growing sense that implicit bias has a complex nature and is underpinned by multiple types of interacting cognitive mechanisms. For example, Holroyd and Sweetman (2016) suggest that implicit biases are heterogeneous and unlikely to be captured, as has traditionally been thought, by a simple distinction between semantic and affective associations. Edouard Machery (2016) argues that implicit biases are traits—dispositions that are exhibited by various socio-cognitive skills (action, perception or decision-making, etc.) in different contexts (see Schwitzgebel (2013) for a related view). Finally, Bryce Huebner (2016) provides a cognitive architecture of implicit bias that involves multiple learning mechanisms calibrated against different aspects of our environment.

Existing implicit bias interventions tend to produce limited effects. Lai et al. (2014) examined seventeen interventions and discovered that only eight of them are effective. A recent meta-analysis study by Forscher et al. (2019) shows that the average effects of the successful interventions are relatively small. Additionally, a review by FitzGerald et al. (2019) also suggests that we need more robust data to determine the effectiveness of interventions. Moreover, most studies analyzed focused on measuring short-term changes with single-session intervention, and it is questionable whether their findings can be generalized to long-term effects. In fact, Lai et al. (2016) show that the eight interventions discussed above does not have effects beyond several hours to several days.[3] Finally,

[1] In this paper, we endorse the widespread view that implicit bias is a mental construct (e.g., an association, attitude, or internal structure) that causes behaviors. However, this view is not unanimously held; for instance, De Houwer (2019) proposes to take implicit bias as a behavioral phenomenon—specifically, behavior that is automatically influenced by cues that function as an indicator of the social group to which one belongs.

[2] There is some disagreement concerning how best to draw the distinction between implicit and explicit attitudes in philosophy and psychology (Brownstein 2018). In the case of implicit and explicit bias, one common way of operationalizing the distinction in scientific practice is to associate them with implicit and explicit measures, respectively. In explicit measures, subjects are asked to report their attitudes in the test, while in implicit measures, their attitudes are inferred from other behaviors (Brownstein 2019). The disagreement will not be the focus of this paper, as we believe it will not affect the arguments of this paper.

[3] See Devine et al. (2012) for a more optimistic result that shows in-person, long-term debiasing can have effects for extended periods of time. However, Forscher et al. (2017) failed to fully replicate the study.

Forscher et al. (2019) show that changes in implicit attitudes do not necessarily translate into changes in behaviors or explicit attitudes.[4]

Here, we suggest that implicit bias's complexity may partly explain the lack of effective interventions. First, current interventions tend to target only a limited number of cognitive mechanisms. Second, current interventions do not consider individual differences in the mechanisms responsible for one's implicit bias. New approaches in interventions need to be explored that can target multiple cognitive mechanisms at the same time and customize them for individuals. In this paper, we systematically examine how AI and related technology can help achieve this end. Finally, before we turn to the next section, it is worth highlighting that there is some ongoing disagreement about how best to intervene on implicit bias. While some researchers focus on its underlying cognitive mechanisms, others (i.e., structuralists) argue that what matter is not cognition but unjust social structures, which should be the main target of intervention (Haslanger 2012). Although this paper focuses primarily on "cognitive intervention," we fully appreciate the importance of structural interventions and will return to them in "Section 6."

3 How AI Has Helped: Existing AI-Assisted Approaches for Bias Reduction

In this section, we propose a framework to evaluate existing approaches for reducing implicit bias. In the next section, we illustrate the utility of our framework for exploring the solution space for promising new approaches. Specifically, we will go beyond algorithmic decisions to discuss alternative ways AI-related technology can facilitate and shape better decisions. To demonstrate how this framework is applied, we use the hiring process as a case study.

3.1 Hiring Process as a Case Study

Our rationale for using the hiring process as a case study is that it involves several phases of decision-making in which common forms of implicit bias can occur. Focusing on the hiring process allows us to consider interventions that target different types of decision-making processes as well as those that encompass the whole hiring process. Another advantage is that relevant bias-reduction technologies have already been developed and are available for investigation. Thus, the focus of our paper will be on the ways in which AI technologies, both existing and prospective, can reduce biased decision-making in the hiring process.

[4] Our paper focuses on AI-assisted intervention on implicit bias rather than on bias in general. Implicit and explicit biases are distinct scientific constructs, and their relation remains a topic of controversy. In addition, it is unclear whether findings in one field can be generalized to the other. For example, a recent study (Forscher et al. 2019) suggests that effective interventions on implicit bias may not always change explicit bias. Finally, implicit and explicit biases bring about different reactive attitudes. For instance, it has been shown that discrimination is considered less blameworthy when it is caused by implicit bias instead of explicit bias (Daumeyer et al. 2019). As a result, we will restrict our discussion to interventions on implicit bias to avoid complicating the discussion. However, the framework we develop in this paper can be adapted to explore intervention on explicit bias.

A hiring process is defined here as a procedure that consists of a series of decisions that are conducted by either an individual agent or a group to select new employees. An idealized hiring process typically consists of four phases. First, during the *pooling phase,* employers create candidate pools using advertisements or by actively reaching out to potential applicants. At the second phase, the *screening phase*, candidates are subject to various types of scrutiny: their resumes are screened, their skills are assessed, and other (cognitively inexpensive) strategies are used to assess their suitability. This screening process reduces a large pool of candidates to a smaller pool for the third phase, the *interview phase*. The interview phase involves face-to-face interaction, skill assessment, and other (more cognitively expensive) ways of selecting the final candidates. Finally, at the *offer phase*, positions are finalized, and successful candidates are presented with a contract, which they may choose to accept, refuse, or negotiate.

3.2 Recent Advances and Ethical Concerns about Artificial Intelligence

AI has an increasing impact on nearly every facet of our lives due to the recent advancement of deep learning and Big Data, which enables deeper integration with other technologies. One of the key capacities of AI that has widely applied is machine learning (ML). ML depends on the development of mathematical theories and algorithms that allow computers to recognize complex patterns in examples or data. For instance, artificial neural networks can be used to train a system to perform object recognition. More powerful methods can also be used to teach AI systems to make connections, hierarchical categorizations, and predictions. AI has significantly increased in power due to deep learning, which enables AI systems to recognize more complex and contextual patterns and solve previously unsolvable problems. Moreover, deep learning can be self-directed; it has the capacity, after some initial set up, to learn continuously as new data arrives without the supervision of engineers. As we will discuss below, these features are useful in real-time interactive AI applications.

Another key reason that AI has become more powerful in the past two decades is the availability of Big Data, i.e., the large amount of diverse and often continuously generated data (Sharda et al. 2020). The development of biometric-related technology has expanded the types of data that AI can draw on. A variety of sensors collect—automatically and in real time—data about such things as a person's heart rate, eye movement, etc. Some of the sensors are backed by AI-related technology. For example, computer vision can automate the processes of acquiring and analyzing visual data by producing meaningful interpretations of objects, faces, or scenes. Natural language processing (NLP), in addition, enables computers to interpret (and generate) written or spoken sentences, as well as translate them from one language/dialect to another. The challenge of processing the vast amount of data available, in turn, is addressed by the advancement of data-related theories and technologies. Data needs to be captured, cleaned, transformed, and analyzed to be useful. Innovation in data science, including the use of ML, has enabled speedy production of quality Big Data.

Ultimately, the availability of quality Big Data, together with deep learning, has led to the advancement of "cognitive computing" (Sharda et al. 2020). Cognitive computing is a type of AI system that is:

(1) Adaptive: it learns in real-time as environments and goals change.
(2) Interactive: humans can interact with it intuitively and naturally.

(3) Iterative: it can identify unsatisfactory solution and conflicting information and request or search for additional information for reprocessing.
(4) Contextual: it solves problems in context-specific ways.

These characteristics allow cognitive computing to improve the quality of the information an AI-powered knowledge-based system can provide for decision support. Moreover, it also enables AI to be further integrated with a host of burgeoning innovations, which, for this reason, we will include in our discussion. For example, augmented reality (AR) can integrate information provided by AI with the user's environment in real-time through visual or auditory equipment, such as Google glasses. Virtual reality (VR) can be combined with AI and related technologies—such as Deepfake, a technique for human image synthesis—to better create a virtual body, character, and environment. Finally, robotics working with visual recognition and NLP can produce robots that perform more complex tasks (either automatically or collaboratively with humans) and better interact with their environment as well as humans. These technologies can enhance the effectiveness of interventions by creating a more natural context in which they can occur. Some existing products for bias reduction have utilized these technologies (see "Sections 3.4, 3.5, and 3.6"). These products, however, have not exploited the full potential of integrating the aforementioned technologies with ML and Big Data (as we will demonstrate in "Section 4").

While our focus in this paper is on discovering novel and effective AI-assisted interventions, it is nevertheless important to address AI-assisted interventions' potential ethical implications. Here, we briefly review some of the pressing ethical concerns that have been discussed in the literature of algorithmic decision.

First of all, *algorithmic bias* and *Big Data bias* are two major challenges (Hajian et al. 2016; Garcia 2016; Suresh and Guttag 2019; Richardson et al. 2019). Algorithmic bias happens when an algorithm produces unfair results (e.g., the disadvantaging of people of color), even if its developers intentionally consider only non-demographic factors in coding (e.g., criminal history). This algorithmic bias happens when non-demographic factors correlate with demographic factors. Big Data bias happens when ML, without the developer's intentions, extracts patterns of prejudice that exist in the collected data. This may result in the relevant prejudice being amplified by the AI system that employs the result of ML. This happens because the Big Data can mirror prejudice existing in human society (e.g., these two challenges confront many existing and promising interventions to be discussed below and are hard to overcome completely). However, they can be ameliorated by adopting ethical guidelines for algorithmic designs and ML,[5] as well as the implementation of fairness-aware data mining and bias correction technology (Hajian et al. 2016; Lu and Li 2012; Obermeyer et al. 2019).

Opacity and *privacy* are also two critical issues that will challenge AI-assisted intervention (Taddeo 2019; Taddeo and Floridi 2018). Opacity refers to the difficulty of clarifying the causal mechanism underlying some algorithmic decisions, and this epistemological difficulty may lead to further complications regarding responsibility and accountability (Castelvecchi 2016; Floridi 2015; Samek et al. 2017; Wachter et al.

[5] For example, Hung and Yen (2020) extract five general principles for protecting basic human rights, including data integrity for reducing bias and inaccuracy through the examination of over 115 principles recently proposed by academies, governments, and NGOs.

2017), which will be discussed in "Section 4.3." Privacy issues happen when poor data security leads to the abuse of data and breaches in privacy, which may result in, say, threats to freedom of expression (Amnesty International UK 2018; Human Rights Watch 2019). As many of the AI-assisted interventions involve ML from personal data, it is unavoidable that they will confront these issues. However, the opacity issue can be ameliorated by developing algorithms for causal explanation or by adopting AI applications that are interpretable—as one of us has argued elsewhere (Hung and Yen 2020; Hung 2020). The privacy issue can also be alleviated by applying the well-developed principles in bioethics (e.g., those about data collection, storage, and reuse) and AI-specific guidelines (e.g., EU, IEEE, and Amnesty International) (IEEE Global Initiative 2016). For example, almost all of these guidelines highlight the principle of data security to ensure that data is under proper protection over its entire life-cycle—which helps reduce privacy breaches. We will point out the relevant ethical issues as we discuss the various types of interventions in the rest of the paper.

3.3 Two Dimensions of the Conceptual Framework

The first dimension (D1) of our framework captures the different types of information AI provides users. In decision-making, it is useful to have information about the current state of affairs (descriptive information), the likelihood of future states (predictive information), and the expected utility of an action (prescriptive information). In keeping with the practice-oriented nature of this paper, we adopt the terms commonly used in knowledge-based systems (KBSs)[6] (Sharda et al. 2020) to label our categories (rows of Table 1).

1) Descriptive analytics: KBSs can consolidate all relevant data in a form that enables appropriate analysis, characterizes data (with descriptive statistics and/or pattern recognition), and visualizes data to inform users of the current state of affairs—as well as informing users of the relevant past and current trends.
2) Predictive analytics: KBSs can provide predictions and inferences about what is likely to happen by analyzing correlative or causal relations among variables and by categorizing cases.
3) Prescriptive analytics: KBSs can provide (recommended) decisions based on what is likely to lead to better outcomes—given the relevant goals (e.g., calculating the expected utility through simulation or optimization models).
4) AI-enhancement without analytics: KBSs can assist interventions without proving analytic information, e.g., by automating the intervention, enhancing other technology involved in the intervention (e.g., robotics), etc.

Some important qualifications are in order. First, the different types of analytics are not completely conceptually independent: predictive analytics depends on descriptive analytics, and prescriptive analytics depends on predictive analytics. Second, there is a sense according to which all interventions are prescriptive, as the KBSs will need to

[6] KBSs are computer programs that generate information to help humans solve problems or generate solutions. AI has played an important role in enhancing the capacity of KBSs by powering knowledge acquisition, representation, and reasoning. In this paper, the use of this term is adopted from the discipline of analytics (Sharda et al. 2020). It is different from the knowledge-based system in AI which represents knowledge and performs inferences explicitly.

Table 1 Framework and existing AI approaches for reducing bias

D 2 D 1	Input-based	Output-based	Cognition-based
Descriptive analytics	➢Descriptive analysis of applicants' demographic information (Eightfold, Entelo, IBM Watson Recruitment)	➢Real-time descriptive analysis of demographic diversity of candidate pool for detecting potentially biased selection decisions at any hiring phases (Eightfold, Entelo)	N/A
Predictive analytics	N/A	➢Predicting the effect of recruiter's behavioral expression, e.g., job advertisement's appeal to potential candidates of different demographic backgrounds (Textio Hire)	➢Predicting and inferring the qualities that make a candidate suitable rather than depending on intuitions rooted in the company's culture and practice, which may be biased (IBM Watson Recruitment, Pymetrics)
Prescriptive analytics	N/A	➢Automated hiring decision: suggesting unbiased evaluative decisions (HireVue, Pymetrics) ➢Automated hiring decision: suggesting unbiased behavioral expression (Textio Hire)	N/A
AI-enhancement without analytics	➢Masking applicants' demographic information, without differentiating whether they induce bias or not (Unbias.io n.d.), Entelo, Blendoor, Eightfold) 1. Removing perceptual cues of implicit bias (Interviewing.io) 2. Creating a virtual space in which candidates can project any avatar they choose (Zaleski 2016) 3. A robotic proxy which allows candidates to control a robot to interact with interviewers (Fair proxy communication) ➢Automatically collecting data of job applicants to reduce implicit bias (Pymetrics, Interviewing.io)	➢Masking recruiters' behavioral expressions, without differentiating biased or unbiased ones 1. Human-free interview (Mya, HireVue), including replacing recruiters with a social robot implemented with standardized questions (Tengai) 2. Creating a virtual space in which recruiters project any avatar they choose (Zaleski 2016)	➢Changing the associations underlying the implicit bias (change-based intervention) 1. Perspective-taking training in VR (Equal reality, Vantage point) 2. Embodying in an avatar with features of the underrepresented in VR (Equal reality) 3. Implement branching narratives in VR to practice making better decisions (Vantage point)

The first dimension (D1) represents different types of analytics playing distinctively crucial roles in the intervention. The second dimension (D2) represents the locus of intervention. In each slot, *arrowheads* indicate types, and *numbers* indicate tokens of interventions

"make a decision" to intervene (perhaps based on some implicit calculation of the expected utility of intervening, which will implicate predictive and prescriptive information as well). However, our category focuses on the type of analytics playing a distinctly crucial role in the intervention itself—because doing so helps conceptualize different existing and promising interventions. Third, our category is characterized at a high-level of abstraction in order to encompass different and more specific ways of producing analytic information. We leave the specifics open because KBSs can in

principle incorporate various different techniques of producing analytics, each of which is appropriate under different contexts.

The second dimension (D2) focuses on the locus of intervention. AI technology can intervene at the different loci of decision-making: at the input, output, and cognition stages of a decision-making process. As a result, interventions are categorized accordingly (columns of Table 1):

(1) *Input-based interventions* reduce implicit bias by managing input information for decision-making. Input information includes perceptual information about an interviewee or the content of an applicant's resume.
(2) *Output-based interventions* manage the output of a biased decision-making process in order to reduce or prevent its harmful effects. Here, examples of output are discriminatory phrases in job ads, unfair evaluative judgments of a resume, sexist speech, and microaggressive behavior[7] toward an applicant, etc.
(3) *Cognition-based interventions* directly target the cognitive processes underlying implicit bias, e.g., training programs that reduce users' biased automatic associations (e.g., of "white" with "good").

By categorizing interventions into these three types, we do not imply that one can think of the input, output, and cognition related to implicit bias independently—rather, they are part of an interactive process. However, by identifying them as distinct loci of intervention in a process, we can better conceive of different possibilities for intervening on bias.

Having introduced the two dimensions of our framework, we will use this framework to assess existing approaches in the following three sections. Currently there are a variety of commercial products as well as proposals for AI-assisted intervention. We systematically evaluate them by categorizing them into types and by placing them within the solution space our framework maps out. Our framework classifies approaches into twelve types (Table 1). Our review will show that eight types of interventions currently exist but require further development. In addition, four types are unexplored and hold potential in the future.

3.4 Existing Input-Based Interventions

We will first review input-based interventions (Table 1, 2nd column). The current approaches take advantage of *descriptive* analytics as well as AI enhancement without analytics. Descriptive, input-based interventions use data visualization and analyze applicants' data to categorize and create labels for candidates' characteristics. The categories or labels include demographic information such as gender, ethnicity, and veteran status (e.g., Eightfold ("Talent Diversity," n.d.), IBM Watson Recruitment (n.d.), Entelo ("Entelo Platform Reports," n.d.)). These interventions provide descriptive information that facilitates the recruiters' understanding of the candidate pool—as

[7] Microaggressions are "brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults toward people of [underrepresented groups]" (Sue et al. 2007, p. 271). Examples include talking over interviewees with a particular demographic background and insensitive comments demeaning interviewee's heritage or identity.

well as facilitating the search for diverse candidates—and can be used to track underrepresented candidates in the hiring process.

There are two types of existing interventions that benefit from AI's enhancement without relying on analytic information. The first, sometimes referred to as *anonymization*, involves masking demographic information about an applicant that can potentially induce implicit biases. It can be implemented at the screening phase, by automating the process of covering up demographic information in resumes. It can also be implemented at the interview phase by employing a robotic proxy that allows candidates to use a robot to interact with interviewers (e.g., Fair proxy communication (Seibt and Vestergaard 2018; Skewes et al. 2019)), or by creating a virtual space in which recruiters and candidates can project an avatar of their own choosing.[8]

The second type of intervention, in contrast, involves automatically collecting data about applicants so that it is less likely to implicate human bias in the data collected. This type of intervention includes automating the evaluation of the candidate's performances and capacities through the KBSs. Candidates may be asked to participate in activities that assess their professional skills (e.g., automated, coding-challenge-based interview provided by Interviewing.io (n.d.)) or their psychological traits (e.g., Pymetrics (n.d.) uses neuroscience-based games to assess candidates' memory capacity, learning skills, speed of reaction, etc.).

However, the above interventions have limitations. Studies show that anonymous recruitment is not always effective in eliminating bias and can create additional disadvantages for underprivileged applicants (Behaghel et al. 2015; Hiscox et al. 2017): removing demographic information prevents recruiters from contextualizing important information embedded in applications. This may result in negative readings of ambiguous signals, e.g., misinterpreting a female candidate's periods of family leave or part-time work as underemployment (Foley and Williamson 2018). Anonymous recruitment may also lead employers to be more influenced by the prejudices induced by other unmasked cues. For example, research suggests that removing criminal histories without masking racial information can increase racial discrimination due to the problematic assumption of racial differences in felony conviction rates (Agan and Starr 2017). Likewise, masking bias-inducing features and responses might take away useful information for interpersonal interaction (e.g., vocal variety and vitality, eye contact, etc.). An applicant's interpersonal skills, for example, can be better assessed via such interaction.[9] Finally, due to the pervasive automation involved in these interventions, ethical issues emerge concerning autonomy, accountability, and responsibility. We will discuss these issues in more detail in "Section 4.3."

[8] Currently, fair proxy communication and interview in virtual space are not products; they are only proposed ideas (Seibt and Vestergaard 2018; Skewes et al. 2019)

[9] Determining which information should be masked to reduce implicit bias is difficult, and the determination needs to be made on a case-by-case basis. In the information technology (IT) industry, for example, the assessment of purely professional skills may be distinguished from other traits related to interpersonal skills (e.g., personality and coordination skills) that may not be essential to the job. So, when evaluating an applicant's coding skills, demographic cues are irrelevant and should be masked. Conversely, in other industries (e.g., insurance sales), masking demographic information could be a loss when assessing the applicant's communication styles that may be essential to the job performance.

3.5 Existing Output-Based Interventions

Output-based approaches utilize descriptive, predictive, and prescriptive analytics, as well as AI-enhancement without analytics (Table 1, 3rd column). The descriptive, output-based interventions currently available detect potentially biased *evaluative decisions* by measuring the demographic diversity of candidate pools in real time. This allows recruiters to identify when candidates from underrepresented groups leave the hiring process and provides room for corrective actions (e.g., Eightfold ("Talent Diversity," n.d.)).

Current predictive output-based interventions provide information about the likely outcome of an expressive behavior. The current approach is to predict the demographic distribution of potential candidates that a job advertisement is likely to attract. For example, the Bias Meter of Textio Hire (Textio n.d.) is a gender tone spectrum that indicates the overall gender bias of job postings. These predictions help a company adjust its job postings to improve hiring diversity at the pooling phase.

The prescriptive, output-based intervention replaces or improves human decision-making with automated decisions from KBSs. First, these systems can evaluate candidates automatically based on data collected from multiple sources in order to produce (recommended) hiring decisions. They do this without the involvement of human recruiters, which reduces the overall influence of recruiter bias on the hiring decisions. For example, some products automatically assess candidates' eligibility and recommend a short-list based on their performance (e.g., Pymetrics (n.d.), HireVue ("CodeVue Offers Powerful New Anti-Cheating Capability in Coding Assessment Tests," 2019)). Also, some systems can produce or suggest unbiased behavioral expressions. For example, Textio Hire can suggest neutral synonyms to replace biased phrases (Textio n.d.).

Finally, there are AI-enhanced (without analytics), output-based interventions that mask recruiter's behavioral expression without differentiating biased from unbiased expressions. The biased behavioral output of recruiters, including biased phrases and microaggressive behaviors (e.g., reduced eye contact with interviewers), can lead to unfair results which are difficult to detect. One example of this is the self-fulfilling influence of social stereotypes on dyadic social interaction (Biggs 2013; Snyder et al. 1977): interviewers can have different styles of interaction based on their stereotypes toward candidates. Their different interaction styles can, in turn, elicit different behaviors from candidates consistent with the interviewers' initial stereotypes, e.g., interviewers' cold and distant treatment toward candidates whom they find less favorable can discourage the candidates from acting in a sociable manner. Automated interviews can mask some of the recruiters' behavioral expressions so as to reduce unfair results (e.g., Mya ("Meet Mya," n.d.), HireVue ("HireVue Video Interviewing Software," n.d.)). Similarly, the technology for masking the bias-inducing cues of applicants can be used to mask recruiters' behavioral expressions as well (e.g., using VR to create avatars for interview (Zaleski 2016)).

However, output-based interventions face limitations too. First, no existing product offers real-time masking of recruiters' microaggressive behaviors at the *interview phase*. Furthermore, even if such intervention is created, it will likely face problems similar to input-based interventions that mask applicant's information: masking (without distinguishing biased or unbiased behavior) may remove too much information

from interpersonal interactions (e.g., eye contact, facial expression) for reliable assessment of communicative skills. Second, human-free interviews have similar limitations. The robot interviewer may not be natural enough because it is automated with fixed scripts and will not be able to assess interpersonal skills.[10] Finally, output-based interventions that rely on automated decision face potential challenges such as algorithmic bias, Big Data bias, and the issue of opacity (see "Section 3.2").

3.6 Existing Cognition-Based Interventions

Cognition-based intervention aims to directly affect agents' cognitive processing (Table 1, 4th column). First, predictive, cognition-based interventions infer the qualities that make a candidate eligible through analyzing the characteristics of the top performers of a given position (e.g., IBM Watson Recruitment Success Score (n.d.), Pymetrics (n.d.)). As such, the qualities for successful candidates are not determined by bias-prone intuitions that are rooted in the company's culture and practice.

Second, existing AI-enhanced (without analytics), cognition-based interventions are change-based. The goal of change-based interventions is to alter the associations underlying implicit biases (Brownstein 2019). For example, taking the perspective of a member of a stereotyped group has been shown to reduce relevant implicit bias (Galinsky and Moskowitz 2000). Some existing interventions utilize VR to create training programs that allow users to adopt the perspective of the underrepresented by embodying an avatar with the relevant features to facilitate bias-reduction (e.g., Equal Reality (n.d.), Vantage point (n.d.)). The other intervention engages users in VR-enhanced real-world scenarios where implicit bias might occur to help users practice making better decisions (e.g., Vantage point (Holpuch and Solon 2018)).

There are some limitations to the effectiveness of these change-based interventions. While Peck et al. (2013) have found that dark-skinned embodiment intervention reduces implicit bias, it is empirically unclear whether this method can be generalized to address other biased factors. In addition, there are worries that the effect of change-based intervention is only small or short term and that changes in implicit associations will not translate into changes in explicit bias or behaviors that maintain intergroup disparities outside of the laboratory setting (Forscher et al. 2019).

In sum, recent advances in AI and related technology have provided us with opportunities for creating a more equitable society. We have examined existing approaches and discussed their limitations. Next, we demonstrate how our framework can help explore new approaches that can overcome some of these limitations.

4 Putting the Framework to Work: New Approaches for Reducing the Harms of Implicit Bias

We begin by illustrating that, by considering D1 (i.e., descriptive, predictive, and prescriptive analytics), we can discover better applications for recent advancements in AI and related technologies, in addition to discovering ways of enhancing human-

[10] Nonetheless, if the robot is too natural, it may trigger the uncanny valley effect—humanoid robots may elicit unintended cold, eerie feelings in human viewers (Mori 1970; MacDorman and Chattopadhyay 2016).

machine and human-human interactions in decision-making. We then show that D2 (input-, output-, and cognitive-based interventions) can help incorporate insights from recent research on implicit bias, cognitive control, and decision augmentation for devising new interventions (see Table 2).

4.1 Utilities of D1: Taking Full Advantage of the Advances in AI and Related Technology

There are at least three general benefits to exploring new approaches using D1.

Exploit the Underutilized Analytics First, distinguishing among existing interventions based on the types of critically-involved analytics allows us to see that existing interventions have underutilized predictive and prescriptive analytics (see Table 1). Our framework suggests that exploiting their full potential will produce better interventions.

Recent developments in deep learning and Big Data have enabled more accurate, quantitative, context-sensitive, and personalized predictions to be produced with a faster speed. The KBSs can generate, in real time, personalized predictions concerning (1) what types of input information causes or correlates with, (2) what sorts of biased evaluative decisions or expressive behaviors will occur, and (3) what types of cognitive processes are implicated in the decision-making (see the 3rd row of Table 2).

AI can generate these predictions with data collected from hiring processes in general (Clabaugh and Matarić 2018), as well as from individualized data acquired in an experimental setting.[11] For instance, it is possible to generate control groups and experimental groups of applicants' resumes and avatars. The two groups would be identical in their relevant qualifications and behaviors but different in their demographic backgrounds. Having set up the groups, data can be collected concerning individual recruiters' biased responses toward applicants, as well as biometric data taken when they review and interact with them. Such data could include information about recruiters' eye-movement to assess the attention they pay to information that could trigger a biased response. This data can also include information about recruiters' body language/facial expressions and spoken/written language in order to gauge any emotional responses that correlate with biased responses. Finally, we can collect information about skin conductance and other physiological data to estimate recruiters' fatigue and stress levels—which often lead to more biased decisions (Clabaugh and Matarić 2018). With this data in hand, ML (assisted by experts) can model the bias patterns of individual recruiters. This knowledge—along with the data collected during, say, actual interviews—can then be used to predict the level of bias in their evaluative decisions, as well as their biased behavioral expressions.

This type of predictive analytics can be extremely useful. Humans are notoriously bad at detecting their own biases (Lai and Banaji 2019). By outsourcing this task to AI—which can alert us when we are likely to be biased—we can refrain from making decisions, or actively

[11] Another example of how AI can help predict human biases is by using ML to detect biases expressed in ordinary language. Caliskan et al. (2017) developed Word-Embedding Association Test (WEAT)—a method of measuring the associations between words. Their model, trained on a corpus of text from the internet, succeeded in replicating the known biases revealed by the Implicit Association Test (e.g., male or female names are associated with career or family respectively). As a result, WEAT can potentially be developed to identify an individual's implicit bias through analyzing the text she produces.

Table 2 Framework showing potential future interventions

D 2 D 1	Input-based	Output-based	Cognition-based
Descriptive analytics	N/A	N/A	N/A
Predictive analytics	➢Predicting the kind of input that would cause biased evaluative decisions and expressive behaviors (See "Section 4.1")	➢Predicting and quantifying the upcoming biased evaluative decisions and behavioral expressions in various phases (See "Section 4.1") ➢Human-machine or human-human collective intelligence (see "Section 4.1")	➢Predicting how a recruiter's level of bias in cognitive processing will be affected by different conditions, e.g. stress, fatigue, etc. (See "Section 4.1"). ➢Potential improvement common to all cognition-based interventions (see "Section 4.2") 1. Adopting an evidence-based approach 2. Customizing/personalizing with better predictive (and prescriptive) analytics 3. Allowing frequent intervention ➢Changing the associations underlying the implicit bias (change-based intervention) 1. Influencing users' multiple cognitive mechanisms through VR enhancement ➢Helping individuals gain better control of the influence of implicit bias on their decision-making and behaviors (control-based intervention) 1. Integrating with AI-related technologies (e.g., AR) to reduce user's cognitive cost, facilitate speedy control, and combine multiple interventions ➢ Enhancing human decision-making capacities via human-machine interactions (augmentation-based intervention) 1. Focusing on natural and complementary human-machine interactions 2. Providing (descriptive, predictive, and prescriptive) information on a need basis
Prescriptive analytics	➢Selectively masking or translating away bias-inducing information about applicants (See "Section 4.1").	➢Selectively masking or translating away demographic information or the biased expressive behaviors of recruiters (See "Section 4.1") ➢Human-machine or human-human collective intelligence (see "Section 4.1")	➢Change-based intervention (see above) ➢Control-based intervention (see above) ➢Augmentation-based intervention (see above)

The first dimension (D1) represents different types of analytics playing distinctively crucial roles in the intervention. The second dimension (D2) represents the locus of intervention

seek out interventions, to reduce bias. Moreover, better predictive analytics can be used to produce better prescriptive analytics and relevant interventions (which we discuss next).

That said, better predictive and prescriptive analytics can come with some cost, including issues of privacy and opacity ("Section 3.2"). More accurate prediction requires more personal data; yet, the collection of personal data comes with the risk of privacy breaches. It is currently an unresolved normative question as to the extent to which a company can legally and ethically collect, store, and use the personal data of candidates and recruiters. Additionally, the manner by which ML generates results is often opaque and inexplicable (Castelvecchi 2016; Wachter et al. 2017). It is unclear whether recruiters should rely on information produced by a "blackbox" algorithm they do not fully understand.

Innovate AI-Related Technologies The second benefit of exploring new approaches using D1 is that it helps consider opportunities that new technologies (e.g., AR, VR, robotics) provide. These technologies, with the help of AI, can enhance the effectiveness of interventions by creating a more natural context under which interventions can occur. Additionally, using data captured by the various sensors, AI can deliver the intervention to the user in a personalized and context-sensitive way. For example, such implementations allow us to selectively "translate" any biased or bias-inducing verbal and visual features/expressions of both applicants and recruiters into neutral features/expressions. This is achieved by combining VR, Deepfake, and NLP with enhanced KBSs that are capable of generating real-time predictions of individual recruiters' biased decisions and expressions (as discussed above) and selectively removing them. This intervention can produce an avatar of an applicant or recruiter in a virtual space that expresses verbal and bodily language with almost identical semantic and emotional content but which includes much less bias-inducing information of applicants or biased expressions of recruiters.

As an input-based intervention, such an approach allows us to selectively "translate away" only the information of applicants predicted to significantly trigger biased evaluation in a particular recruiter while at the same time retaining information that does not trigger such biased evaluations. One advantage is that the input-based intervention will enable the recruiter to take advantage of the remaining demographic information to properly contextualize applicants' performance and reach fairer evaluative decisions. Moreover, the intervention will allow the recruiter to better interact with applicants and facilitate proper evaluations of their relevant interpersonal skills. Both features ameliorate the limitations of existing input-based interventions that mask all the demographic information of applicants.

As an output-based intervention, this approach can selectively "translate" only a recruiter's significantly problematic expressive behaviors, again striking a balance between bias-reduction and the allowance of social interactions in order to assess relevant skill sets. As a result, this approach can overcome a key limitation of existing output-based interventions that automatize interviews or mask all of the recruiters' behavior non-selectively.

However, this new technology may invite attendant harm. Translating away interviewers' biased expressions means that interviewees lose the opportunity to recognize that their interviewers are biased and to choose whether to address the relevant issue on the spot—for example, addressing mansplaining by asserting one's epistemic authority.

This limitation may thus result in worse outcomes for the interviewee overall.[12] Moreover, by taking away important information for decision-making, it also invites ethical concerns such as AI paternalism (i.e., AI increases a human's own good at the cost of restricting their autonomy). We will discuss this issue in "Section 4.3."

Enhance Human-Machine Interactions Finally, D1 can help us explore four enhanced modes of human-machine interactions. These interactions are enhanced to the extent that AI systems, with enhanced predictive and prescriptive analytics, improve their capacity to:

1) Collaborate with humans in discovering solutions to complex problems. Recent AI's continuous and fast learning from Big Data has enabled it to interact with humans in real time, provide context-appropriate support, and complement humans' strengths and weaknesses.
2) Assist interpersonal interactions in decision-making, such as improving the quality of interpersonal communications and collective decision-making.
3) Train us to make better decisions by shaping our cognitive process in a personalized, naturalistic, and effective training environment.
4) Automatically make context-appropriate decisions for complex problems.

Consider, for example, new interventions that employ human-machine group collaborations. They are made possible based on predictive analytics that quantify individual recruiters' reliability (e.g., the degree to which they exhibit bias) in their evaluation. Aggregating evaluative decisions, under the right conditions, can lead to more reliable collective decisions than those made by individuals—this phenomenon is called the Wisdom of the Crowd Effect (Surowiecki 2005). One good way of aggregating individual decisions is to do so after weighing them by their reliability (as predictive analytics provided by KBSs), which further enhances the reliability of collective decision-making. Moreover, AI-agents using a variety of algorithms to make automated hiring decisions (prescriptive analytics) can be included as recruiters as well. This can result in a further improvement in reliability when a group of decision-makers have different backgrounds—such that their judgments reflect independently generated and divergent points of view. This new mode of intervention allows recruiters to reach less biased decisions by complementing one another's strengths and weaknesses.

However, integrating automated decision into a collective decision-making framework not only raises ethical issues of algorithmic bias, Big Data bias, and the issue of opacity ("Section 3.2"), but it also raises new issues about individual and *collective* responsibility. We shall discuss these emerging issues in "Section 4.3."

4.2 Utilities of D2: Incorporating Insights from Recent Empirical Research

As we have shown in "Section 3," D2 illuminates the fact that cognition-based interventions have been under-explored by existing approaches. However, there is a

[12] A possible solution to this attendant harm focuses on reducing the implicit bias of interviewers. Since AI detects bias, it can also be programmed to alert the interviewers for correction while masking the biased expressions to the interviewees. The detection record can be used by senior managers to choose better interviewers.

rich cognitive scientific literature that we can draw on to design more effective AI-assisted, cognition-based interventions. There are at least two further types of cognition-based interventions available in the literature. The first is control-based intervention, which aims to help individuals gain better control of the cognitive processes underlying implicit bias and to prevent the processes from affecting their decisions and behaviors. The second is augmentation-based intervention, which enhances human decision-making capacities via human-machine interactions in which a computer acts as a companion or advisor in an ongoing context-sensitive way. In what follows—and illustrating the utility of our framework—we consider these promising new changed-based, control-based, and augmentation-based interventions.[13] Moreover, we will show that by considering D1 and D2 together, we can address some worries raised in "Section 3.6" concerning the ineffectiveness of cognition-based interventions.

Advance Change-Based Interventions with Evidence-Based Approach, Personalization, and AI-Related Technology First, the recent literature on implicit bias can help advance better AI-assisted, change-based interventions. For example, research shows that only a select set of change-based interventions have robust short-term effects, e.g., competition with shifted group boundaries, shifting group affiliations under threat, etc. (Lai et al. 2016). However, none of the existing approaches take advantage of these findings. There is also a rich literature concerning ways to enhance the effectiveness of change-based interventions (Brownstein 2019). By incorporating these recent findings, we will be more likely to develop effective AI-assisted interventions.

Moreover, we can improve the effectiveness of change-based interventions by reflecting on D1 and D2 together. For example, we can maximize the potential of these various interventions by personalizing cognition-based interventions. With the help of better analytics, KBSs can identify the types of implicit bias that require the most attention—as well as the most effective interventions—for particular recruiters. For instance, KBSs may determine, with predictive analytics, that a recruiter is relatively more biased against women of color in contexts of evaluating intelligence (Madva and Brownstein 2018). KBSs may then implement the interventions to target the relevant biased associations in a more focused way. Additionally, KBSs can determine what types of change-based interventions will work better for the recruiter by running predictive analysis on the feedback collected. Such analytic information can help implement interventions that produce the most benefit with limited resources.

Third, we can also exploit new, AI-related technology and new modes of human-machine interaction to improve the interventions' effectiveness. For example, VR can (1) create a vivid and rich virtual social and physical environment for reducing biases, in which (2) the users can engage more actively in an immersive and self-directed way, in order to (3) influence multiple cognitive mechanisms (including visual, auditory, cognitive, emotional, evaluative, etc.). All of these features have been shown to promote more effective changes in one's implicit attitudes (Byrd 2019). Moreover, VR can also make interventions more fun by turning them into a VR game or other entertainment experience (e.g., using Deepfake technology to give any Hollywood

[13] However, we should not think of the three types of cognition-based interventions as a final and unrevisable category of cognition-based intervention. This is because as our knowledge about the mechanisms of implicit bias grows, new types of cognition-based intervention may become available.

movie an all-Asian cast to increase positive experiences with outgroup members). As the availability of VR equipment approximates smartphones, change-based interventions will no longer be restricted to lab settings. Rather, interventions can be undergone daily for an extended period. This has the potential to increase their long-term effectiveness.[14]

Facilitate Control-Based Interventions with Analytics and Automation Another example of how D2 helps explore promising AI-assisted approaches is drawing our attention to control-based interventions. Empirical research has suggested that some control-based interventions may be efficacious. Among these are implementation intentions, which are "if-then" plans that specify a response that a decision-maker can perform upon encountering a particular perceptual cue. For example, if I see a dark-skinned face, then I will respond by thinking "good" (Gollwitzer 1999). Compared with change-based approaches, control-based interventions may lead to immediate behavioral change through self-control, and there is also evidence suggesting that these interventions have long-term effects (Lai and Banaji 2019; Burns et al. 2017; Monteith et al. 2013). Research suggests, in other words, that AI-assisted control-based interventions are worth exploring.

Again, bringing D1 into consideration can be beneficial. One potential criticism of control-based interventions is that they may not be practically feasible because they may place great demands on cognitive resources—in particular, they tax an agent's scarce resources for self-control (Botvinick and Braver 2015). For example, implementation intention requires subjects to be on the lookout for the specific "if" condition and recall the relevant "then" condition to control their behavior. It thus requires considerable effort to implement just one implementation intention—much less than the number required to adequately address bias in the hiring process. AI-related technology, such as AR, can make control-based interventions more feasible by taking the cognitive burden off the users. For example, a pair of Google glasses can help detect several different "if" conditions in the environment and remind the user of the relevant "then" conditions. Moreover, better predictive analytics can further enhance the quality of interventions. A real-time prediction can help initiate the implementation intention either before or shortly after the "if" condition obtains. A faster prediction entails a more effective control-based intervention because cognitive control is most effective when control-related signals are generated early enough to have an impact upon biased decision-making. Finally, given that the large number of potential "if" conditions in the environment may still overwhelm the user despite automation, prescriptive analytics provided by KBSs can help determine which "if" conditions should be prioritized in the relevant contexts.

Note that while the above change-based and control-based interventions come with clear benefits, both require significant time and resources to train recruiters who are participating in the hiring processes (although see Madva (2017) for an argument that such commitment may not be as big as has been assumed). Besides, these interventions clearly invite ethical concerns related to privacy, as they require the collection of massive amounts of personal

[14] However, we need to be careful of the unforeseen ethical consequences of interventions (such as those involving VR). For example, Madary and Metzinger (2016) point out that VR can induce illusions of embodiment and change one's long-term psychological states. Risky content and privacy are critical issues too. Therefore, they offer a list of ethical recommendations as a framework for future study. While there will always be unforeseeable risks involved in new technology, such research will help us minimize it.

data (See "Section 3.2."). Additionally, as some of the control-based interventions involve manipulation without the user's consent (e.g., nudging), they can violate the user's autonomy—beyond the ethical concern of AI paternalism (see "Section 4.3" for more discussion). These drawbacks may pose an obstacle for companies or recruiters who wish to adopt this strategy to reduce bias.

Take Full Advantage of Augmentation-Based Interventions One final case illustrates the benefits of bringing together all aspects of our framework when exploring potential AI-assisted interventions. D2 draws our attention to the relative neglect of decision augmentation, a field that examines how human-machine interaction can enhance the quality of decision-making (Jarrahi 2018). Although this entire paper can be seen as an application of decision augmentation, we have not emphasized the field's key insights. To begin with, this field stresses natural human-machine interaction and hence focuses on systems that could engage with humans using natural language as well as intuitive data visualization. Also, it focuses on the complementarity between humans and machines. For instance, when tackling a problem that is difficult for humans to solve, humans may seek help from AI to analyze Big Data. AI can also provide feedback in a user-friendly form such as a narrative explanation that summarizes complex data in a narrative form. Finally, humans can ultimately accept or reject such advice after taking on board broader considerations that may be hard for AI to take on board.

For example, a social robot can work as an advisor to the recruitment team during interviews, similar to a moral advisor in the case of moral enhancement (Savulescu and Maslen 2015). It can do so by incorporating both the existing and potential interventions discussed above. For example, it can provide "translated" information about interviewees for proper contextualization (e.g., by interpreting their performance relative to their access to opportunity) at the appropriate time. Moreover, it can lead the team to engage in deliberation that is less likely to be biased (e.g., by adopting more criteria-based judgments using criteria that track actual performance). It can also bring attention to potentially biased responses in interviews. There are newer strategies that can be incorporated as well. It is possible for social robots, for example, to create a more inclusive interview environment by discouraging sexist speech with a subtle disapproving frown (Paiva et al. 2018). Moreover, the robot could also act as a "Socratic Assistant" to provide empirical support, improve conceptual clarity and argumentative logic, etc. (Lara and Deckers 2019). Of course, as augmentation-based intervention can potentially integrate all interventions discussed above, it will confront all the challenging ethical issues for each type of intervention.

In short, D2 can help us systematically explore new AI-assisted interventions by bringing together insights from newly emerging empirical research and AI-related technologies. Overall, we have shown that our framework is useful for developing new approaches for reducing the harm of implicit bias.[15]

[15] The interventions proposed in this paper are generally based on currently available AI and AI-related technologies; however, their advancement relies on the development of AI research in some domains. In particular, predictive interventions face the challenge of modeling and predicting the behavior of an individual accurately; on top of that, prescriptive interventions, in order to suggest decisions to its user, require a causal model, which represents how the intervention leads to results for a particular user (Albrecht and Stone 2018; Sheridan 2016). Finally, we need empirical research to validate the effectiveness of the specific implementation of these interventions.

4.3 Emerging Ethical Issues for Promising Interventions

The new approaches we discussed can remedy some of the limitations that face existing interventions, but they also raise new ethical challenges, including the difficulty of attributing individual and collective responsibility—as well as the threat to human autonomy. We will not be able to resolve these controversial issues in this paper; however, we aim to show that they are not insurmountable problems that prevent us from adopting these promising AI-assisted interventions.

First, the attribution of moral responsibility is complicated by automated decision-making by KBSs (Doshi-Velez and Kortz 2017), especially in contexts of collective decision-making involving a group of human and AI agents (Winsberg et al. 2014). To handle the issues of the attribution of individual responsibility, one promising way is to adopt ethical guidelines that require humans to be the ultimate decision-makers in decisions involving KBSs (Hung and Yen 2020). So an individual (e.g., a manager) needs to make an explicit decision to transfer some power of decision-making to KBSs, ensuring that it is human agent who is ultimately responsible for the decisions.

Moreover, Miller's (2017, 2018) account of *collective moral responsibility* can also help the attribution of collective responsibility. According to this account, agents with different *roles* in the collective decision-making process can have a collective end in a chain of responsibility (i.e., each agent makes a different and distinct contribution, according to their roles, to the collective end and shares collective responsibility). So when a recruiter makes a morally wrong decision based on a problematic recommendation by a KBS, which in turn results from the negligence of a software engineer, both the recruiter and the engineer are collectively responsible and accountable (praised or blamed) for the wrong decision.[16] In short, existing theoretical frameworks about collective responsibility can help hold the right agent responsible and accountable for the wrong decision and hence alleviate the ethical concern.

Second, as AI-assisted interventions shape human decisions through interfering with the deliberation processes, violation to human autonomy (i.e., roughly, the freedom of self-determination and self-control) can become a serious ethical concern. For example, when AI increases a human's own good at the cost of restricting autonomy, AI paternalism may happen. Likewise, nudging, which manipulates decision-making without consent or understanding on behalf of the individuals involved, may also violate their autonomy. Again, these issues are difficult but not completely unsolvable. With regard to AI paternalism, introducing the well-developed guidelines from bioethics (e.g., opting-out, informed consent, and the principle of autonomy) can be helpful. For example, the principle of autonomy could be helpful (Amnesty International UK 2018; Floridi and Cowls 2019; Anonymous, forthcoming). According to the principle, (i.e., respect for the rights of self-determination), one should determine by herself whether to exchange partial autonomy (e.g., determining which route to go) for some good (e.g., the convenience of trip planning on Google maps), thus preventing KBSs from undermining her autonomy.

[16] According to Miller (2018), *responsibility* is about the ability to fulfill a duty, and *accountability* is about the liability to respond to one's performance of duties. Accountability presumes responsibility, but is not identical with it. Please see Miller (2018) for further distinction of the two notions.

About the worry that nudging may violate autonomy, Barton (2013) argues that in some cases, nudging (e.g., tobacco health warnings) can in fact foster autonomy (e.g., helping smokers to control themselves better). It has also been argued by Engelen and Nys (2020) that such a worry may be overblown and should be reassessed by clarifying the notion of autonomy. According to them, nudging's threat to autonomy is rarely supported by a proper understanding of autonomy. Moreover, given a graded understanding of autonomy, nudging can restrict one's autonomy without completely violating it.[17] In short, this account helps alleviate the ethical concern of nudging, even if it does not clear it away completely. To summarize, the promising AI-assisted approaches discussed above are ethically viable ways of addressing the problem of implicit bias; however, further research into the ethical implications of these approaches still needs to be pursued.

5 Designing a Better Hiring Process with the Framework

The previous discussions have mainly focused on individual interventions, the majority of which target just one individual recruiter at a specific phase of the hiring process. However, the hiring process can involve multiple recruiters, the biases of whom can affect any phase of the recruitment process. As such, ensuring a fair hiring process means paying attention to the hiring process as a whole. This implies that intervention design should aim at providing a multi-factorial approach that combines interventions to restructure the hiring processes. Again, our framework can work as a useful conceptual tool here: it helps design interventions that work synergistically by clarifying each approach's function, its locus of intervention, as well its strengths and weaknesses.

Different approaches (AI-assisted or not) are not mutually exclusive; rather, they can often complement each other to enhance the overall efficacy of the intervention. For example, existing intervention strategies heavily rely on masking the demographic contents of resumes. While it has positively influenced the recruitment of people from certain underrepresented groups (e.g., women; Krause et al. 2012), this practice may also disadvantage candidates of lower socioeconomic status by obscuring the fact that their achievements are exemplary relative to the relatively limited opportunities they have had (see "Section 3.4"). This problem can be addressed by replacing this problematic resume screening process with a new form of low-cost interview, thus restructuring the hiring process. This lost-cost interview method reduces implicit bias by combining an input-based intervention that collects alternative data from candidates during an automated interview, with an output-based intervention that makes fully automated evaluative decisions.

It is also possible to combine control-based interventions with the practice of selectively masking applicants' demographic information. On the one hand, control-based interventions can be overwhelming if the recruiters are constantly alerted with cues for control during their decision. On the other hand, masking demographic

[17] Engelen and Nys (2020) propose the concept of *perimeters of autonomy*, according to which changes in an agent's options within the perimeters can occur without precluding his autonomy because he still has a range of options to choose from. Nonetheless, there may be an issue about how to draw the perimeters.

information may lead to inadequate contextualization of the applicant's behaviors and performance as discussed previously. By combining the two interventions in the hiring process, masking can reduce the frequency of cues for control (as some of the triggers for implicit bias are masked), while control-based interventions, such as implementation intention, can reduce biased decision-making based on unmasked, contextualizing, yet potentially bias-inducing information.

This design strategy has implications for research on implicit bias intervention. Implicit bias, as many scholars have emphasized, involves multiple interacting cognitive mechanisms. The strategy of combining multiple complementary interventions has the benefit of simultaneously targeting a multiplicity of underlying cognitive processes. Doing so may result in more effective and sustained changes which overcome a key problem identified by current literature: the failure of individual cognitive interventions to produce long-term effects.

6 Conclusion and Future Directions

To summarize, implicit bias is a complex problem underpinned by multiple, interacting cognitive mechanisms. We have proposed a framework to assess existing AI-assisted interventions, explore future approaches, and restructure hiring processes. We are confident that the framework can be applied to tackle implicit bias in domains other than job recruitment, such as policing and healthcare. Granted, there are unresolved limitations facing individual interventions, some of which generalize to many other AI applications—including, but not limited to, the normative issues discussed in "Section 3.2" and "Section 4.3." However, we are optimistic that future research will lead to the development of technological and social solutions that address them appropriately.

While we have focused exclusively on interventions that target cognitive mechanisms, structuralists may argue that our framework fails to address structural problems. However, recent research has stressed the dynamic interactions between cognitive and structural factors. Soon (2019), for example, has emphasized the dynamic causal processes by which biased mind and structure sustain themselves mutually. Liao and Huebner (2020) also argue that implicit bias is a multifaceted phenomenon involving dynamic interaction and mutual dependence among cognitive, social, and physical factors.[18] That is, "individualistic interventions can have structural effects, and vice versa" (Soon 2019, p. 3), and they are equally important in achieving equity (Saul 2018; Zheng 2018). For example, a cognitive intervention can draw a company's attention to low inclusivity in its policies, as well as any micro-aggressive behaviors in its workplace. This could lead to institutional change within the company, which in

[18] The complex interaction between cognitive and structural factors can have unpredictable consequences. It is exemplified in the change of implicit and explicit antigay bias before and after same-sex marriage legalization. Ofosu et al. (2019) found that implicit and explicit antigay bias decreased before the legalization of same-sex marriage. Nevertheless, the change of attitude following legalization differs depending on whether the legalization was passed locally: a deeper decrease was found if the legalization was passed locally, whereas an increase following federal legalization in states that never passed local legalization. However, note that Tankard and Paluck (2017) found that federal legalization led individuals to change their perceptions of social norms regarding gay marriage, but not their personal attitudes.

turn makes it more likely to adopt a more comprehensive framework of cognitive interventions. In fact, we believe, as our framework suggests, that interventions which target multiple cognitive mechanisms, and interact dynamically with the unjust social and physical environments in which they are embedded, have the most potential to affect positive individual and structural changes.

Acknowledgements For helpful discussions and feedback on earlier drafts of this work, thanks to Michael S. Brownstein, Acer Chang, Caitrin Donovan, Ivan Gonzalez-Cabrera, Julia Haas, Richard Heersmink, Bryce Huebner, Calvin Lai, Eric Schwitzgebel, Jacob Sparks, and two anonymous referees.

Funding information This work is supported in part by an Academia Sinica Fellowship to Dr. Linus Ta-Lun Huang, sponsored by Academia Sinica, Taiwan. This research is also funded in part by the Ministry of Science and Technology Taiwan to Dr. Tzu-wei Hung (MOST 107-2410-H-001-101-MY3).

References

Agan, A., & Starr, S. (2017). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics, 133*, 191–235.

Albrecht, S. V., & Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence, 258*, 66–95. https://doi.org/10.1016/j.artint.2018.01.002.

Amnesty International United Kingdom. (2018). Trapped in the matrix: Secrecy, stigma, and bias in the Met's gangs database. https://reurl.cc/8lmnzy. .

Barton, A. (2013). How tobacco health warnings can Foster autonomy. *Public Health Ethics, 6*(2), 207–219.

Behaghel, L., Crepon, B., & Le Barbanchon, T. (2015). Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics, 7*, 1–27.

Biggs, M. (2013). *Prophecy, self-fulfilling/self-defeating. Encyclopedia of Philosophy and the Social Sciences*. Inc: SAGE Publications. https://doi.org/10.4135/9781452276052.n292. isbn:9781412986892.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control. *Annual Review of Psychology, 66*(1), 83–113.

Brownstein, M. (2018). *The implicit mind: Cognitive architecture, the self, and ethics*. New York, NY: Oxford University Press.

Brownstein, M. (2019). Implicit bias. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019).

Burns, D., Parker, M., & Monteith, J. (2017). Self-regulation strategies for combating prejudice. In C. Sibley & F. Barlow (Eds.), *The Cambridge Handbook of the Psychology of Prejudice* (pp. 500–518).

Byrd, N. (2019). What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese.*

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*, 183–186.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*(7623), 20–23. https://doi.org/10.1038/538020a.

Chamorro-Premuzic, Tomas (2019). Will AI reduce gender bias in hiring? Harvard Business Review.

Clabaugh, C., & Matarić, M. (2018). Robots for the people, by the people. *Science Robotics, 3*(21).

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, 84, 103812.

De Houwer, J. (2019). Implicit bias is behavior: A functional-cognitive perspective on implicit bias. *Perspectives on Psychological Science, 14*(5), 835–840.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48*(6), 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003.

Doshi-Velez, F., & Kortz, M. (2017). Accountability of AI under the law: The role of explanation. In *Berkman Klein center working group on explanation and the law*. Berkman Klein: Center for Internet & Society working paper.

Dunham, C. R., & Leupold, C. (2020). Third generation discrimination: An empirical analysis of judicial decision making in gender discrimination litigation. *DePaul J. for Soc. Just, 13*.

Eightfold AI. (n.d). Talent Diversity. Retrieved from https://reurl.cc/EKp05m

Engelen, B., & Nys, T. (2020). Nudging and autonomy: Analyzing and alleviating the worries. *Review of Philosophy and Psychology, 11*(1), 137–156.

Entelo. (n.d.). Entelo Platform Reports. Retrieved from https://reurl.cc/Gko62y

Equal Reality. (n.d.). Retrieved from https://equalreality.com/index

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: A systematic review. *BMC Psychology, 7*(1), 29. https://doi.org/10.1186/s40359-019-0299-7.

Floridi, L. (2015). *The ethics of information*. Oxford University Press.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*.

Foley, M., & Williamson, S. (2018). Does anonymising job applications reduce gender bias? Understanding managers' perspectives. *Gender in Management, 33*(8), 623–635. https://doi.org/10.1108/GM-03-2018-0037.

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of Experimental Social Psychology, 72*, 133–146.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of change in implicit bias. *Journal of Personality and Social Psychology, 117*, 522–559.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*(4), 708.

Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal, 33*(4), 111–117.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126).

Haslanger, S. (2012). *Resisting reality*. Oxford: OUP.

HireVue. (2019). CodeVue offers powerful new anti-cheating capability in coding assessment tests. Retrieved from https://reurl.cc/24D9An

HireVue. (n.d.). HireVue video interviewing software. Retrieved from https://reurl.cc/NapMKk

Hiscox, M. J., Oliver, T., Ridgway, M., Arcos-Holzinger, L., Warren, A., & Willis, A. (2017). Going blind to see more clearly: Unconscious bias in Australian public service shortlisting processes. *Behavioural Economics Team of the Australian Government.* https://doi.org/10.1016/j.jmrt.2015.05.003.

Hodson, G., Dovidio, F., & Gaertner, L. (2002). Processes in racial discrimination. *Personality and Social Psychology Bulletin, 28*(4), 460–471.

Holpuch, A., & Solon, O. (2018, May 1). Can VR teach us how to deal with sexual harassment? In *The Guardian* Retrieved from https://reurl.cc/A1KreQ.

Holroyd, J., & Sweetman, J. (2016). The heterogeneity of implicit biases. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and philosophy, volume 1: Metaphysics and epistemology*. Oxford University Press.

Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 1). Oxford: Oxford University Press.

Human Rights Watch. (2019). *World report*, 2019 https://reurl.cc/6g641d. .

Hung, T.-w. (2020). A preliminary study of normative issues of AI prediction. *EurAmerica, 50*(2), 205–227.

Hung, T.-w. & Yen, Chun-pin (2020). On the person-based predictive policing of AI. *Ethics and Information Technology.* https://doi.org/10.1007/s10676-020-09539-x.

IBM Knowledge Center (n.d.). Retrieved from https://reurl.cc/W4k9DO

IEEE Global Initiative. (2016). *Ethically aligned design. IEEE Standards*, v1.

Interviewing.io. (n.d.) Retrieved from https://interviewing.io/

Jarrahi, M. (2018). Artificial intelligence and the future of work. *Business Horizons, 61*(4), 577–586.

Krause, A., Rinne, U., & Zimmermann, K. (2012). Anonymous job applications in Europe. *IZA Journal of European Labor Studies, 1*(1), 5.

Lai, C. K., & Banaji, M. (2019). The psychology of implicit intergroup bias and the prospect of change. In D. Allen & R. Somanathan (Eds.), *Difference without domination: Pursuing justice in diverse democracies*. Chicago, IL: University of Chicago Press.

Lai, C. K., Marini, M., Lehr, A., Cerruti, C., Shin, L., Joy-Gaba, A., et al. (2014). Reducing implicit racial preferences I. *Journal of Experimental Psychology: General, 143*(4), 1765.

Lai, C. K., Skinner, L., Cooley, E., Murrar, S., Brauer, M., Devos, T., et al. (2016). Reducing implicit racial preferences II. *Journal of Experimental Psychology: General, 145*(8), 1001.

Lara, F., & Deckers, J. (2019). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics.* https://doi.org/10.1007/s12152-019-09401-y.

Liao, S., & Huebner, B. (2020). Oppressive Things. Philosophy and Phenomenological Research. https://doi.org/10.1111/phpr.12701.

Lu, J., & Li, D. (2012). Bias correction in a small sample from big data. *IEEE Transactions on Knowledge and Data Engineering, 25*(11), 2658–2663.

MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition., 146*, 190–205.

Machery, E. (2016). De-freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy*, *Metaphysics and epistemology* (Vol. 1, pp. 104–129). Oxford: Oxford University Press.

Madary, M. & Metzinger, T.K. (2016). Real virtuality: A code of ethical conduct. Recommendations for good scientific practice and the consumers of VR-technology. *Front. Robot. AI* 3:3. https://doi.org/10.3389/frobt.2016.00003.

Madva, A. (2017). Biased against debiasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice. *Ergo*, 4.

Madva, A., & Brownstein, M. (2018). Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Noûs, 52*(3), 611–644.

Miller, S. (2017). Institutional responsibility. In M. Jankovic & K. Ludwig (Eds.), *The Routledge handbook of collective intentionality* (pp. 338–348). New York: Routledge.

Miller, S. (2018). *Dual use science and technology, ethics and weapons of mass destruction*. Springer.

Monteith, J., Woodcock, A., & Lybarger, E. (2013). *Automaticity and control in stereotyping and prejudice*. Oxford: OUP.

Mori, M. (1970/2012). The uncanny valley (K. F. MacDorman & N. Kageki, trans.). IEEE Robotics and Automation, 19(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811.

Mya. (n.d.). Meet Mya. Retrieved from https://mya.com/meetmya

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

Ofosu, E. K., Chambers, M. K., Chen, J. M., & Hehman, E. (2019). Same-sex marriage legalization associated with reduced implicit and explicit antigay bias. *Proceedings of the National Academy of Sciences*, 116, 8846-8851.

Paiva, A., Santos, P., & Santos, F. (2018). Engineering pro-sociality with autonomous agents. *Proc of AAAI*.

Peck, T., Seinfeld, S., Aglioti, S., & Slater, M. (2013). Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition, 22*(3), 779–787.

Pymetrics. (n.d.). Retrieved from https://www.pymetrics.com

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., & Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour*, 1–9.

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review, 94*, 192–233.

Samek, W., Wiegand, T., & Muller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU journal: ICT Discoveries, 1*.

Saul, J. (2018). Should we tell implicit bias stories? *Disputatio., 10*(50), 217–244.

Savulescu, J., & Maslen, H. (2015). Moral enhancement and artificial intelligence. *Beyond Artificial Intelligence* (pp. 79–95). In J. Romportl, E. Zackova, J. Kelemen (eds), Beyond artificial intelligence. Springer.

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New essays on belief*. New York: Palgrave Macmillan.

Seibt, J., & Vestergaard, C. (2018). Fair proxy communication. *Research Ideas and Outcomes, 4*, e31827.

Sharda, R., Delen, D., & Turban, E. (2020). Analytics, data science, & artificial intelligence: Systems for decision support. Pearson.

Sheridan, T. B. (2016). Human–robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 58*(4), 525–532. https://doi.org/10.1177/0018720816644364.

Skewes, J., Amodio, D., & Seibt, J. (2019). Social robotics and the modulation of social perception and bias. *Philosophical Transactions of the Royal Society B, 374*(1771).

Snyder, M., Tanke, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology, 35*, 655–666.

Soon, V. (2019). Implicit bias and social schema. *Philosophical Studies*, 1–21.

Sue, D., Capodilupo, C., Torino, G., Bucceri, J., Holder, A., Nadal, K., & Esquilin, M. (2007). Racial microaggressions in everyday life. *American Psychologist, 62*(4), 271.
Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv* preprint arXiv:1901.10002.
Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Anchor Books.
Sweeney, L. (2013). Discrimination in online ad delivery. *Queue, 11*(3).
Taddeo, M. (2019). Three ethical challenges of applications of artificial intelligence in cybersecurity. *Minds and Machines, 29*(2), 187–191.
Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.
Tankard, M. E., & Paluck, E. L. (2017). The effect of a supreme court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science, 28*, 1334–1344.
Textio. (n.d.). Textio hire. Retrieved from https://textio.com/products/
Unbias.io. (n.d.) Retrieved from https://unbias.io/
Vantage Point. (n.d.). Retrieved from https://www.tryvantagepoint.com/
Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics.
Winsberg, E., Huebner, B., & Kukla, R. (2014). Accountability and values in radically collaborative research. *Studies in History and Philosophy of Science Part A, 46*, 16–23.
Zaleski, Katharine. (2016). Virtual reality could be a solution to sexism in tech. Retrieved from https://reurl.cc/vnezZk
Zheng, R. (2018). Bias, structure, and injustice: A reply to Haslanger. *Feminist Philosophy Quarterly, 4*(1).

Affiliations

Ying-Tung Lin[1] · Tzu-Wei Hung[2] · Linus Ta-Lun Huang[2,3]

[1] Institute of Philosophy of Mind and Cognition, National Yang-Ming University, No.155, Sec.2, Linong Street, Taipei 112, Taiwan

[2] Institute of European and American Studies, Academia Sinica, No. 128, Sec. 2, Academia Rd., Nankang District, Taipei 115, Taiwan

[3] Department of Philosophy, University of California, 9500 Gilman Drive # 0119, La Jolla, San Diego, CA 92093-0119, USA