

A Literature Review of New Direction in Implicit and Explicit Stereotypes Researches

Yicheng Zhu

Qibao High School, Shanghai, 201102, China

ZhuYu@microport.com

Abstract. Implicit stereotypes are first noticed in the 1980s, nearly half century later than explicit ones. Since then, an array of indirect methods such as Implicit Association Test (IAT) and Attention Misattribution Procedure (AMP) were invented to further analyze this subtle subject. Recently, with the development of Internet and AI, researchers have generally combined these emerging techniques with conventional study on explicit and implicit stereotypes. The current review mainly introduces three new directions of stereotype study and analyzes a basket of representative and comprehensive researches, as examples, to propose the main strengths and weaknesses of these techniques. First, stereotype tests via Internet have strengths of large samples, improved accuracy, and raising public awareness, but they suffered from non-random recruitment and unstable judgment. Second, applying word embedding to stereotype study has strengths of improved efficiency and comprehensiveness, as well as tracing back history, but it suffered from limitation as book texts and limited samples. Third, utilizing Large Language Models (LLMs) to stereotype study is yet a new direction with little published researches, but the current review cautions that the divergence between human thoughts and LLMs' outputs may matter. In conclusion, the current review postulated that these three directions have a promising future and can deeply inspire further study.

Keywords: Stereotypes, attitudes, PI: International, word embedding, LLMs.

1. Introduction

Study on stereotypes can date back to 1930s and had significant influence on the improvement on social equality and the reduction of bias and discrimination. Both explicit and implicit stereotypes are vital portions of individuals' attitudes. For decades, with growing sense of equality, explicit stereotypes have been significantly and deliberately erased across nations and cultures. However, it was not until 1980s that researchers began to apply the Dual Process Theory to researches of stereotypes, and noticed the existence of implicit stereotypes and their potential influence [1]. Implicit stereotypes, in comparison with explicit ones, are relatively harder to alter and thus still influencing individuals' opinions and behaviors subtly.

Unlike words like "discrimination", "prejudice", or "bias", it is notable that the word "stereotype" has a less negative valence. In fact, in *Stereotypes as Explanations*, authors concluded three mainstream guiding principles to identify stereotypes, which were (a) stereotypes are aids to explanation, (b) stereotypes are energy-saving devices, and (c) stereotypes are shared group beliefs [2]. In conclusion, though stereotypes may lead to problems such as discrimination, they are more commonly considered as helpful paradigm and formation of beliefs. Thus, instead of simply discarding stereotypes, researchers have tried figuring out how they work, how they change, and, most importantly, to what extent they will influence participants' minds and behaviors.

Nowadays, there is an emerging study direction of applying cutting-edge computer science to psychology studies, to help researchers glean and scan massive samples, or to act as virtual substitutes for human participants. This current trend also influenced studies on explicit and implicit stereotypes. In combination with new computer techniques such as word embedding and LLMs, the conventional methods like IAT and AMP can now refresh their scope and depth, yielding more comprehensive and extensive results [3, 4]. In the current literature review, three emerging directions in stereotype studies, as well as their unique strengths and weaknesses, will be mainly discussed and introduced, then, according to the current trends, I will provide general predictions on future research direction.

2. Tests via Internet: The Project Implicit

2.1. An Introduction of the Project Implicit: International

The conventional researches, first focusing on explicit stereotypes then taking implicit stereotypes into account, were inevitably constricted by time and space. They mainly focused on either time or space variations and thus could be roughly classified into four categories: (a) the combined study of both explicit and implicit stereotypes; (b) the study of cross-cultural divergences in stereotypes; (c) the study of time variation in explicit stereotypes; and (d) the study of the intersection of these previous traits. Previous papers showed that researchers had long troubled by the limited samples and effortful collecting process, thus initiated to establish websites, first in the United States as Project Implicit: US, then globally as PI: International, to “facilitate comprehensive studies across multiple cultures and multiple years” [5, 6]. Currently, the datasets have significantly raised the public’s awareness of implicit stereotypes, and have delivered and developed vital samples for multiple researches (see <https://www.projectimplicit.net/resources/publications/>) in various subjects such as multiform stereotypes, bias-reduction, and relationship between bias [5].

2.2. Strengths

2.2.1 Large samples

It is notable that the PI: International was perhaps the largest dataset currently focusing on stereotypes and attitudes. It collected 2,386,123 samples in trials about 6 bias attitudes and 1 gender stereotype. Across 11 years, all the samples were collected from 34 countries (Canada and Switzerland had bilingual data, so 36 samples in total), thus yielding 252 datasets. The large scope of samples helped corroborate many conclusions in previous papers in cross-cultural and comprehensive dimensions, as well as held up unique conclusions that were divergent to, or even opposite to, previous researches. For instance, in the male-science/female-arts stereotype, only 6 out of 36 cultural groups (i.e. 36 sample groups) showed the expected conclusion (male respondents revealing higher implicit gender-science stereotypes than female respondents), while 17 countries showed the opposite result and 13 countries showed no overall disparity [6]. Such across-culture conclusion could only be unveiled by such extensive and comprehensive study across multiple countries.

2.2.2 Improved accuracy

Previous work propounded various method to preclude disturbance and to distinguish implicit and explicit stereotypes. Here, as a comprehensive work, the PI: International adopted multiple and relatively accurate methods. The project used Implicit Association Test to measure indirect (implicit) stereotypes and attitudes and used self-report 7-point Likert items and 11-point feeling thermometers to measure direct (explicit) stereotypes and attitudes (see <https://www.projectimplicit.net/>) [3]. Additionally, demographic information was collected before all the trials. Since implicit and explicit stereotypes and attitudes are known to be separate but related, these methods enabled further comparison between the two series of data and made the studies more reliable and comprehensive.

Besides, researchers excluded all the incomplete IAT trials and fast trials (<300ms) in advance to exclude possibly indeliberate or casual trials and adopted formations to reduce possible non-representativeness in certain countries [6]. Whether respondents chose to start tests through the official website or not, they would be re-oriented to the test websites that were written in respondents’ native language to ensure no delays in their responsive time, which might lead to errors since IAT trials gauged the participants’ implicit stereotypes based on their difference in responsive time between congruent and incongruent statements.

2.2.3 Raising awareness

Implicit attitudes and stereotypes were new concepts, and, previously, much of the public had not noticed or been aware of this potential cause of bias and prejudices [5]. Unlike previous researches in limited areas, with PI: International’s unique advantage of wide spreading over conventional

laboratory trials, it played a vital role in the raising of the public's awareness. It potentially reminded the public that even the behavior of those people harboring no or little explicit bias might be inevitably influenced by implicit stereotypes. This finding helped reduce inequality and bias in societal and political levels.

2.3. Weaknesses

2.3.1 Non-random samples

Though it was efficient and convenient, this innovative method of recruiting respondents might predictably lead to some limitations. According to the respondents' demography, the majority of respondents were (a) told to visit the website by instructors or mentors or (b) self-directed or recommended to the website [6]. This led to diverged proportion and emphasis across different social groups. For example, the participants skewed to be "more young, liberal, and female than the population," and they are more likely to be people who are interested in psychology or relative issues [7].

Besides, the numbers of participants and samples diverged across countries and were hugely influenced by the countries' Internet access. And the respondents in the low internet-access countries were more likely to be "more affluent, urban, or educated respondents" [6]. As a result, the representativeness of data in different countries and cultures might largely vary. Though some modifications have been applied to the dataset, researchers still reminded the users of interpreting it cautiously, and innovatively utilizing methods to synthetically correct the needed samples.

The PI: International aimed to establishing a global sample resource, but currently, it was far from providing global-coverage. In Africa particularly, only data in South Africa were collected. The researchers of PI: International are now planning to extend the websites to more countries in their native language. This method then shows direction for future explicit and implicit stereotype studies in a societal level, which is a potentially vital resources of samples, across place and time, for future studies.

2.3.2 Unstable judgment

The online trials aimed at providing a "virtual laboratory" by using every participant's computer or laptop. Thus, there was inevitable variability in the test environment and inter-person performance, making the judgments of participants less stable than traditional pencil-and-paper trials. All the trials were completed in an open environment, using keyboards to quickly respond to pictures and words, which might lead to inevitable disturbance and distraction. Besides, IAT trials demanded the participants to correct their mistakes and would automatically exclude this certain trial. Participants might hesitate if they had made too many mistakes in previous trials, which might prolong their future responsive time. Notably, after how many errors they would slow down and how long they would re-think might differ between persons, which potentially caused errors.

3. Analyzing Stereotype Changes via Diachronic Texts

3.1. An Introduction of Word Embedding

Though massive samples might contribute to overcoming the geographically restrictions, it was much harder to overcome the limitation of time. Since the concept of have begun to form in the 20th century and implicit stereotypes were noticed even later, in the 1980s, the existing researchers are far from cover all stereotype changes in history. Therefore, instead of focusing on human participants, researchers began to notice an indirect method: analyzing diachronic bias and stereotypes through texts in previous books, which could partly reflect the attitudes of their authors and the social trends in certain era [8].

Innovations in the field of Natural Language Processing (NLP) made this assumption possible, by applying emerging word embedding techniques to transform words into vectors, then computing the geometric relationship (mean average cosine similarity, MAC) between certain groups of vectors.

The MAC represented to what extent the target words associated with other words semantically and lay foundation for further insights. Notably, the stereotypes represented in text was a combination of both explicit and implicit stereotypes according to the dual process theory, thus laying foundations for further studies [1]. The reliability of this indirect method, has been corroborated by researches' capturing certain societal shifts and its known groups validity [9, 10].

3.2. Strengths

3.2.1 Improved efficiency and comprehensiveness

The adoption of word embedding technique to process natural language in psychology studies largely facilitated the current study and inspired future researches. This innovation enabled researchers to examine a large scope of texts during a very vast period of time and largely reduced the time required to scan or analyze the texts, which improved efficiency and comprehensiveness in studies. In fact, thanks to the efficiency, researchers could collect text samples from certain authors' books or from massive online corpora like Google Books [11, 12].

This unique advantage over conventional methods gave rise to an array of new conclusions. Many of them were impossible to obtain from previous study and some of them might even be opposite to previous conclusions. For example, previous studies showed that traits related to groups changed across time and their mean valence were generally getting positive [13, 14]. But in Charlesworth, T. E. S.'s paper, with a continuous data resource instead of intermittent ones from a cohort of previous studies, researchers noticed that only the top10 or 50 traits changed [12]. If they extend their scope to a list of nearly 14,000 words and over 600 traits, some traits and stereotypes were stable across decades and their mean valence, contrast to previous conclusions, remained stable or became slightly more negative.

3.2.2 Tracing back in history

It is notable that the majority of the previous inter-country or inter-time study only focused on explicit stereotypes since implicit stereotypes were not noticed until the 1990s [1]. Using word embedding techniques, however, researchers were able to trace back the existed stereotypes in history and the extension of time was largely determined by the scope of datasets and corpora. For example, Garg's study collaborated with the US Census to collect 100y of text data and Charlesworth's study, applied Google Books corpus to collect 200y of text data [9, 12]. This advantage enabled researchers to cast light on the stereotypes in history and figure out how different kinds of stereotypes changed across time differently. For instance, the researchers found that, in history, stereotypes relevant to inborn traits (e.g. nationality and race) were more open to change, while stereotypes relevant to social status and health (e.g. age, weight, and class) were relatively more stubborn. Some groups of people like "the overweighted" were linked to various stereotypes across time, since the social trends and fashions had altered (they were estimated to be affluent and well-cared in the past, but to be undisciplined and lazy in modern society) [12].

3.3. Weaknesses

3.3.1 Limitations in book text

One inevitable limitation of the new method was that the samples were from book texts, which meant that there was a lag time between the composition and the publication and renowned books might be reprinted or cited. The indirect book texts also made researchers barely possible to identify words' precise meanings and valence in certain era. For instance, words like "gay" and "broadcast" had their meaning drifted due to either the emerging concept or the emerging technology [15]. What is more, for words like "awful" and "awesome", both their semantic meaning and valence changed across time. Since the word embedding technique could only process words in their current mainstream meaning, the drift might cause errors.

3.3.2 Limited samples

Though the history corpora largely manifest in the mainstream opinion in certain times, it still has inevitable limitations. Nowadays, large corpora and datasets mainly focused on Eng-all texts, unconsciously excluded other culture from the samples. Besides, since the study was based on written books, it only represented the author social group, a biased slice of society, and was expected to be well-educated and wealthier than the population. Thus, the existing studied were far from depicting the global societal stereotypes in certain eras. Further works will be needed to expand the scope of datasets and corpora thus delivering more general data to future studies on societal stereotypes.

4. Promising Future in LLMs & Stereotypes

4.1. An Introduction of LLMs in Psychology

Recent years have witnessed the robust development of Large Language Models (LLMs) and many innovative researchers have already applied them to psychology studies in different fields [16]. In social and cultural psychology particularly, researchers have long been troubled by ethical approvals and recruiting participants, so researchers began to try LLMs as virtual participants to replace at least some of the human participants [17]. Previous studies showed that LLMs characterize human attitudes and they even have personality and values, in substitute for human participants [18, 19]. Studies have found that LLMs can make biased decisions or behaviors, and are now trying to eliminate them to reduce negative bias and stereotypes in society and researches [20]. Talboy et al., asserted that LLMs may preserve or even exaggerate existing human bias in their output [21]. These fundings corroborated that stereotypes in LLMs exist and lay foundation for future application. In the future studies on explicit and implicit stereotypes, LLMs are predicted to function in at least two directions. First, like NLP, they are capable of being scientific tools, helping researchers analyze large textual data, model social phenomenon, and propound unique insights [22]. Next, they can mimic human responses and behaviors, in substitute for human participants [23].

4.2. Strengths

No researches have applied LLMs to explicit and implicit stereotype studies in scale yet, so it is difficult and too bold to identify their strengths and weaknesses currently. But at least some of the advantages can be predicted: as efficient scientific tools and substitutes for human participants, LLMs can largely reduce the cost and conducting time of future studies, thus promoting more comprehensive and extensive researches.

4.3. Weaknesses

The application of LLMs in scales will at least lead to the following three problems. First, though intelligent they seemed to be, their output should not be misinterpreted as the presence of human-like thoughts, but instead viewed as a complex matching process modeled to bubble up the subsequent words [24]. Second, unlike human participants whose thinking mechanisms were generally known through previous empirical studies, the internal working mechanisms of LLMs remain a black box in cognitive level. Thus, their competence in different aspects may vary and cause potential risks. Finally, policy and norms about the LLMs are far from established, to what extent they can replace human participants and whether they can or should do that remain unsettled. Further questions such as AI morals and abuses need to be further discussed.

5. Conclusion

Recent years have witnessed the rapid development in new technologies, which significantly altered the study on psychology. In the current review, three new directions of study on explicit and

implicit stereotypes and attitudes, and their strengths and weaknesses in comparison with previous study are mainly discussed.

First, the current review introduces the well-known dataset Project Implicit: International, which conducts tests on both explicit and implicit stereotypes via internet in the respondents' native language. The project has advantage in its unprecedentedly large samples across time and space; its improved accuracy and comprehensiveness, as a result of previous iteration; and its unique contribution to the propaganda of the concept of implicit stereotypes. However, it also suffers from weaknesses such as its non-random method of recruiting respondents and the instability due to the "virtual laboratory" environment.

Second, the current review discusses the emerging word embedding technique, an excellent combination of computer science and psychology. It enables researchers to transform natural language into vectors, then efficiently analyze large text samples in corpora or datasets and conduct further study. In comparison with conventional methods, it benefits from its improved efficiency and comprehensiveness and the unique capability of tracing back history. However, the analyzing of existing book texts also brings weaknesses such as the limitations of books and limited sample resources.

Finally, the current review pays attention to LLMs, a popular technique nowadays. The previous study showed the feasibility and reliability of applying LLMs to stereotype study as substitutes for human respondents and scientific assistants. Though not many researches have been conducted in this direction, the current review cautions that the outputs of LLMs should not be misinterpreted as kinds of thoughts and further effort on their working mechanisms, morals, and norms are needed before LLMs are widely accepted and applied.

In conclusion, these three directions discussed above are generally promising, for they not only yielded influential results but only inspired future study. However, future works are needed in these directions, like expanding the PI: International to more countries and cultures; enlarging the corpora and datasets to better utilize word embedding techniques; and further applying LLMs to study on stereotypes. In addition, the relative researches can serve as unique foresights into the relationship between human minds and AI, a nebulous and unsettled question needed to be discussed.

References

- [1] Greenwald, A. G. (1990). What cognitive representations underlie social attitudes? *Bulletin of the Psychonomic Society*, 28 (3), 254–260. <https://doi.org/10.3758/bf03334018>.
- [2] McGarty, C., Yzerbyt, V. Y., & Spears, R. (Eds.). (2002). *Stereotypes as explanations: The formation of meaningful beliefs about social groups*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511489877>.
- [3] Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 (6), 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>.
- [4] Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005b). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89 (3), 277–293. <https://doi.org/10.1037/0022-3514.89.3.277>.
- [5] Ratliff, K. A., & Smith, C. T. (2021). Lessons from two decades of Project Implicit. In Krosnick, J. A., Stark, T. H., & Scott, A. L. (Eds.). *The Cambridge handbook of implicit bias and racism*. Cambridge, England: Cambridge University Press.
- [6] Charlesworth, T. E. S., Navon, M., Rabinovich, Y., Lofaro, N., & Kurdi, B. (2023). The project implicit international dataset: Measuring implicit and explicit social group attitudes and stereotypes across 34 countries (2009-2019). *Behavior research methods*, 55 (3), 1413–1440. <https://doi.org/10.3758/s13428-022-01851-2>.
- [7] Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of experimental psychology. General*, 148 (6), 1022–1040. <https://doi.org/10.1037/xge0000623>.

- [8] Nerbonne, J. (2013). The secret life of pronouns. What our words say about us. *Literary and Linguistic Computing*, 29 (1), 139–142. <https://doi.org/10.1093/llc/fqt006>.
- [9] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115 (16). <https://doi.org/10.1073/pnas.1720347115>.
- [10] Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the “known-groups” method. *Applied Psychological Measurement*, 8 (3), 295–305. <https://journals.sagepub.com/doi/10.1177/014662168400800306>.
- [11] Burrows, J.F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
- [12] Charlesworth, T. E. S., Caliskan, A., & Banaji, M. R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences of the United States of America*, 119 (28), e2121798119. <https://doi.org/10.1073/pnas.2121798119>.
- [13] Bergsieker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology*, 102 (6), 1214–1238. <https://doi.org/10.1037/a0027717>.
- [14] Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton Trilogy revisited. *Personality and Social Psychology Bulletin*, 21 (11), 1139–1150. <https://doi.org/10.1177/01461672952111002>.
- [15] Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1605.09096>.
- [16] Ke, L., Tong, S., Chen, P., & Peng, K. (2024). Exploring the frontiers of llms in psychological applications: A comprehensive review. <https://doi.org/10.48550/arXiv.2401.01519>.
- [17] Gray, K., Yam, K. C., Zhen’An, A. E., Wilbanks, D., & Waytz, A. (2023). The psychology of robots and artificial intelligence. *The handbook of social psychology*. Dillion, D., Tandon, N., Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7), 597-600. <https://doi.org/10.1016/j.tics.2023.04.008>.
- [18] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2022). Out of One, Many: Using Language Models to Simulate Human Samples. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2209.06899>.
- [19] Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An Exploration of Personality, Values and Demographics. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2209.14338>.
- [20] Abramski, K., Citraro, S., Lombardi, L., Rossetti, G., & Stella, M. (2023). Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. *Big Data and Cognitive Computing*, 7 (3). <https://doi.org/10.3390/bdcc7030124>.
- [21] Talboy, A. N., & Fuller, E. (2023). Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2304.01358>.
- [22] Salah, M., Al Halbusi, H., & Abdelfattah, F. (2023). May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research. *Computers in Human Behavior: Artificial Humans*, 1 (2). <https://doi.org/10.1016/j.chbah.2023.100006>.
- [23] Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, 380 (6650), 1108-1109. <https://doi.org/10.1126/science.adi1778>.
- [24] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30 (4), 681-694. <https://doi.org/10.1007/s11023-020-09548-1>.