

Vertrauen ist gut, Kontrolle ist besser: Experimentelle Philosophie und die Qualität empirischer Daten

THOMAS PÖLZLER

Institut für Philosophie, Universität Graz, Österreich

Stellen Sie sich vor, Sie fragen einen Bekannten nach seiner Meinung zum neuen österreichischen Arbeitszeitgesetz. „Bedenklich“, antwortet er in festem Tonfall. Sie nicken zustimmend und lehnen sich zurück. Einen Moment später jedoch realisieren Sie, dass Ihr Bekannter gar nicht von seinem Handy aufgesehen hat. Seine Antwort hat er überdies schon gegeben, noch bevor Sie Ihre Frage ganz ausgesprochen hatten. Und als Sie ihn nun bitten, seine Meinung zu begründen, bemerkt er nur, dass das Arbeitszeitgesetz im Juni eingebracht worden sei. – Sie werden unsicher. Ist Ihr Bekannter *wirklich* der Ansicht, dass das Gesetz bedenklich ist?

Schon seit längerem beschäftige ich mich mit empirischen Zugängen zu philosophischen Fragen. Ein Stipendium des Österreichischen Wissenschaftsfonds ermöglicht mir gegenwärtig, zum ersten Mal auch selbst an entsprechenden Studien mitzuwirken. Gemeinsam mit Jen Wright (einer Psychologin am College of Charleston) untersuche ich die metaethischen Intuitionen gewöhnlicher Leute. Insbesondere interessiert uns, ob diese Leute an die Existenz objektiver moralischer Wahrheiten glauben, d.h., ob sie die moralischen Eigenschaften von Handlungen als unabhängig von individuellen und gesellschaftlich geteilten moralischen Überzeugungen betrachten (Pölzler 2018; Pölzler und Wright in Begutachtung).

Eine meiner überraschendsten Erkenntnisse im Rahmen dieser interdisziplinären Zusammenarbeit betrifft die Qualität der von meiner Kollegin und mir erhobenen empirischen Daten. Auf unsere erste Pilotstudie erhielten wir zahlreiche Antworten, die mir ein wenig wie jene des unaufmerksamen Bekannten aus dem obigen Beispiel erschienen. Sie mochten echte

Überzeugungen reflektieren – oder eben auch nicht. In der Folge setzte ich mich daher etwas eingehender mit dem Problem unzureichend guter empirischer Daten auseinander. Dieser Blog-Beitrag skizziert meine Haltung zu diesem Problem, speziell im Kontext der experimentellen Philosophie.

Gute Daten definieren

Um die Qualität von Datensätzen einschätzen und verbessern zu können, muss man eine Vorstellung davon haben, was „gute“ Daten sind. Im Folgenden geht es mir um Güte in einem minimalen Sinn. Mit guten Daten meine ich Daten, die aus einer hinreichenden Anstrengung der Probanden/innen hervorgingen (unabhängig davon, ob die Daten relevant für die Einschätzung der zur Debatte stehenden Hypothese sind, sich replizieren lassen, usw.). Im Englischen wird dies als „sufficient effort responding“ bezeichnet (Huang et al. 2012; Huang et al. 2015).

Philosophisch relevante empirische Studien haben zumeist die Intuitionen von gewöhnlichen Leuten zum Gegenstand. In diesem Zusammenhang erfordert hinreichende Anstrengung – und folglich Güte – meines Erachtens vor allem dreierlei. Erstens, die Probanden/innen müssen die Anweisungen, Fragen und Antworten der Studie aufmerksam lesen. Zweitens, sie müssen sich über ihre Intuition zur vorliegenden Frage klarzuwerden versuchen. Und drittens, sie müssen sich bemühen, auf jene Weise zu antworten, die diese Intuition am besten reflektiert.

Eine geeignete Stichprobe wählen

Wie hoch die Qualität der Daten ist, die eine Studie zu Tage fördert, hängt unter anderem von ihrer Stichprobe ab. Traditionellerweise wurden psychologische und sozialwissenschaftliche Untersuchungen an Studierenden durchgeführt. Seit einigen Jahren greift man jedoch, auch in der experimentellen Philosophie, immer stärker auf Online-Plattformen wie Amazon Mechanical Turk zurück. Über solche Plattformen gelangen Forscher/innen schnell, unkompliziert und relativ kostengünstig an Daten. Bei den Probanden/innen kann es sich überdies auch um ältere und nicht-„seltsame“ Personen handeln (im Sinne von Personen, die nicht aus „Western, Educated, Industrialized, Rich and Democratic (WEIRD) societies“ stammen) (vgl. Henrich et al. 2010).

Amazon Mechanical Turk Arbeiter/innen nehmen als Erwerbsquelle an Hunderten, mitunter sogar Tausenden Umfragen pro Jahr teil. Mit Hinblick auf ihre Anstrengungen wird

dies oft als kontraproduktiv erachtet. Man stellt sie sich als strikte Zeit- und Einkommens-Maximierer/innen vor. Tatsächlich legen Untersuchungen jedoch nahe, dass MTurker/innen im Durchschnitt mindestens ebenso aufmerksam, wenn nicht sogar aufmerksamer als Studierende sind (Bartneck et al. 2015; Hauser und Schwartz 2015; Paolacci et al., 2010). Auch unsere Studie zu metaethischen Intuitionen hat diese Einschätzung bestätigt. Gemessen an den weiter unten dargelegten Kriterien Antwort-Geschwindigkeit, Verständnisfragen und offene Fragen, waren unsere MTurk-Daten jenen von Studierenden deutlich überlegen.

Die Qualität der Daten überprüfen

Inwieweit sich Probanden/innen beim Absolvieren einer Studie anstrengen, lässt sich nicht direkt beobachten. Es gibt jedoch zahlreiche Indikatoren dafür (Huang et al. 2012; Huang et al. 2015). In unserer oben erwähnten Pilotstudie zogen Prof. Wright und ich etwa folgende Kriterien zur Überprüfung der Datenqualität heran:

- *Antwort-Geschwindigkeit*: Wie lange ließ sich der/die Proband/in für die Beantwortung einzelner Fragen und der Umfrage im Gesamten Zeit?
- *Verständnistests*: Wie viele der Verständnisfragen, die wir nach bestimmten Anweisungen und Fragen stellten, beantwortete der/die Proband/in korrekt?
- *Offene Fragen*: Wie relevant und Missverständnis-frei waren die Erklärungen des/der Probanden/in, um die wir ihn/sie nach der Beantwortung bestimmter Fragen baten?

Das Ergebnis unserer Analyse war ernüchternd (selbst im Fall der MTurk Stichprobe). Um sich unsere Anweisungen, Fragen und Antworten aufmerksam durchzulesen, hätte ein/e Proband/in etwa 40 Minuten benötigt. Die Studie wurde jedoch im Durchschnitt schon nach etwa 30 Minuten abgeschlossen; mitunter sogar in weniger als 20 Minuten. Sowohl unsere Verständnistests als auch unsere offenen Fragen offenbarten zahlreiche Missverständnisse. Vereinzelt wurden Erklärungen aus Online-Quellen wie Wikipedia in die Antwortfelder kopiert. Es schien sogar, dass manche Proband/innen nicht einmal der englischen Sprache mächtig waren.

Die Qualität der Daten verbessern

Schon vor der Datenanalyse kann dem Problem des „insufficient effort responding“ bis zu einem gewissen Grad entgegengewirkt werden. Für unsere Hauptstudie hoben Prof. Wright und ich z.B. die Bezahlung der MTurk Probanden/innen und die Anzahl ihrer für die Teilnahme erforderlichen positiven Bewertungen an. Außerdem wiesen wir darauf hin, dass

die Umfrage Verständnis- und Aufmerksamkeitstests enthält, dass man sie nicht im Eiltempo absolvieren kann und dass wir nur an gewissenhaften und ehrlichen Antworten interessiert sind.

Probanden/innen, die trotz solcher Maßnahmen nicht die nötige Anstrengung aufbringen, sollten meines Erachtens in vielen Fällen von der Analyse ausgeschlossen werden. Auf Grundlage welcher Kriterien dies am besten getan wird, hängt u.a. von der Art der Studie ab. Auch potentielle Nachteile von Maßnahmen zur Überprüfung und Verbesserung der Datenqualität müssen berücksichtigt werden. Aufmerksamkeits-Tests z.B. erlauben zwar Rückschlüsse auf das Anstrengungs-Level der Probanden/innen. Sie können diese aber auch in Richtung bestimmter (für sozial wünschenswert erachteter) Antwortoptionen lenken und sich negativ auf die Aufmerksamkeitsleistung im Rest der Studie auswirken (Vanette 2017). Auch kann die Nicht-Berücksichtigung nachlässiger Probanden/innen die Repräsentativität einer Stichprobe schmälern (Berinsky et al. 2013).

Nach sorgfältiger Abwägung aller Vor- und Nachteile ergänzten Prof. Wright und ich unsere Hauptstudie um eine Reihe spezifischer Aufmerksamkeitstests. Zusätzlich überprüften wir die Datenqualität auf Grundlage der im letzten Abschnitt genannten Indikatoren. Nur wenn ein/e Proband/in hinreichend aufmerksam war, sich beim Ausfüllen der Studie hinreichend Zeit ließ, hinreichend viele Verständnisfragen korrekt beantwortete und hinreichend relevante und Missverständnis-freie Erklärungen gab, berücksichtigten wir seine/ihre Antworten bei der Auswertung der Daten. Von ursprünglich 172 Probanden/innen erfüllten 117 diese Bedingung.

Eine Empfehlung

Als erfahrene Psychologin überraschte und beunruhigte meine Kollegin die schlechte Qualität unserer Pilotstudien-Daten weniger als mich. Auch Gespräche mit anderen Forschern/innen legen nahe, dass dem Problem zu wenig Beachtung geschenkt werden könnte. Häufig wird den Probanden/innen entweder blind vertraut oder ein gewisses Maß an anstrengungslosem Antworten als letztlich irrelevant betrachtet. Dabei kann eine solche Tendenz gerade im Kontext von philosophisch relevanten Untersuchungen (wo es oft um sehr feine Unterscheidungen geht) Ergebnisse deutlich verzerren. Schon ein kleiner Prozentsatz nachlässiger Probanden/innen reicht theoretisch hin, um eine bestehende Korrelationen zu verschleiern oder eine nicht bestehende Korrelationen vorzutäuschen (Huang et al. 2015).

Denken Sie noch einmal zurück an den eingangs geschilderten Fall. Eine Person äußerte sich zum neuen österreichischen Arbeitszeitgesetz. Nachdem diese Äußerung ein unzureichendes Maß an Anstrengung zu reflektieren schien, zogen wir ihren Informationsgehalt in Zweifel – und das schien ganz natürlich. Meines Erachtens sollten auch experimentelle Philosophen/innen (und Psychologen/innen und Sozialwissenschaftler/innen im Allgemeineren) in ihren Studien stärker auf diese Weise vorgehen. Anstrengung muss gefördert werden. Anstrengung muss überprüft werden. Und wenn sie unter einem bestimmten vorher festgelegten Level bleibt, sollten die entsprechenden Datensätze in vielen Fällen verworfen werden. Kurz gesagt: „Vertrauen ist gut, Kontrolle ist besser“.

PS.: Unsere Studie hat schließlich überraschenderweise ergeben, dass die meisten Personen die Existenz objektiver moralischer Wahrheiten leugnen. Insbesondere betrachten sie solche Wahrheiten als abhängig von ihren eigenen moralischen Überzeugungen und den in ihrer Gemeinschaft vorherrschenden Überzeugungen. Dieses Resultat ist unseres Erachtens auch philosophisch relevant. Es wirft Zweifel an einem weit verbreiteten Argument zu Gunsten des moralischen Objektivismus auf (dem Argument, dass die meisten Personen an die Existenz objektiver moralischer Wahrheiten glauben und deshalb Anti-Objektivisten die Beweislast tragen).

Bibliografie

- Bartneck, C.; Duenser, A.; Moltchanova, E.; Zawieska, K. (2015): Comparing the similarity of responses received from studies in amazon's mechanical turk to studies conducted online and with direct recruitment. *PloS one* 10 (4).
- Berinsky, A. J.; Margolis, M. F.; Sances, M. W. (2014): Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58 (3), 739-753.
- Hauser, D. J.; Schwarz, N. (2015): Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* 48 (1), 400–407.
- Henrich, J.; Heine, S. J.; Norenzayan, A. (2010): The weirdest people in the world. *Behavioral and Brain Sciences* 33 (2-3), 61-83.
- Huang, J. L.; Curran, P. G.; Keeney, J.; Poposki, E. M.; DeShon, R. P. (2012): Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology* 27 (1), 99-114.
- Huang, J. L.; Bowling, N. A.; Liu, M.; Li, Y. (2015): Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology* 30 (2), 299-311.
- Paolacci, G.; Chandler, J.; Ipeirotis, P. G. (2010): Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 411-419.

Pölzler, T. (2018): How to measure moral realism. *Review of Philosophy & Psychology*.

Pölzler, T.; Wright, J. C. (in Begutachtung): Anti-realist pluralism: A new approach to folk metaethics.

Vannette, D. L. (2017): *Using attention checks in your surveys may harm data quality*.

<https://www.qualtrics.com/blog/using-attention-checks-in-your-surveys-may-harm-data-quality/>