meta(φ)

Werte und Normen

Der Wert der Fälschung
• Nicola Grossrieder

Pragmatistische Wahrheit und induktives Schliessen
• Micha Herrmann

A Defense of Perspectivism about Ought Against the Argument from Advice
• David Lussi

Overdemandingness and Supererogation
• Timo Junger

Intergenerational Distribution
• Dela Wälti

Genocide: Essentialism and Identity
• Sarah Heinzmann

Affordances and the Normativity of Emotions
• Rebekka Hufendiek

Was ist eine Frau?
• Rahel Wehrlin

Herausgeberin
Fachschaft Philosophie der Universität Bern
Länggassstrasse 49a, 3012 Bern, Schweiz
mit Unterstützung des Instituts für Philosophie der Universität Bern

Redaktion
meta(φ)
Fachschaft Philosophie, Universität Bern
metaphi.philo@lists.unibe.ch

Chefredaktion
Vera Moser

Lektorat
Timo Junger, Isabel Käser, Vera Moser, Xenia Schmidli, Andrea Schnyder

Produktion und Gestaltung
Aleksandar Nikolić

Coverbild
Timo Junger

Druck
Kromer Print AG Lenzburg

ISSN
Gedruckte Ausg.: meta(φ) ISSN:2297-9948
Online-Ausg.: meta(φ) [Elektronische Ressource] ISSN:2297-9956

Text-Einreichungen
Einreichungen zur Veröffentlichung sind erwünscht an metaphi.philo@lists.unibe.ch. Die Auswahl der Texte erfolgt themenunabhängig, jedoch wird philosophische Relevanz des Themas vorausgesetzt. Für jede Ausgabe akzeptieren wir bis zu zwei Einreichungen pro Person. Es besteht kein Anrecht auf Veröffentlichung, weder bei erfolgreicher Einreichung noch nach Anfrage durch ein Redaktions-Mitglied. In diesem Magazin veröffentlichte Texte gelten als wissenschaftliche Publikation und sind somit zitierfähig.

Spenden
Die Produktion des meta(φ) ist kostspielig. Sofern Ihnen das Journal gefällt, freut sich die Fachschaft Philosophie daher sehr über finanzielle Unterstützung. Spenden werden folgender Empfängerin zuteil.

Fachschaft Philosophie
Universität Bern, 3012 Bern
IBAN: CH53 0900 0000 3050 3703 2
Zahlungsgrund: «metaphi»

Erste Worte

Liebe Leser*innen

In einer Welt, wo sich die unterschiedlichsten Menschen versammeln, sich vielfältige Werte entwickeln und diverse Normen entstehen, ist es nicht immer leicht, sich selbst und die eigenen Erkenntnisse einzuordnen. Doch erst wenn wir unsere Erkenntnisse ordnen, können wir bestimmen, welche Normen wir predigen und welche Werte wir umsetzen sollten. Ordnen können wir sie aber erst dann, wenn wir sie haben.

Die vorliegende Ausgabe des meta(φ) vereint acht philosophische Texte unter dem Titel *Werte und Normen*. Im ersten Teil widmen sich Nicola Grossrieder und Micha Herrmann dem Thema *Wert und Erkenntnis* und beschäftigen sich mit den Werten von Wahrheit und Falschheit. Grossrieder macht den Auftakt und zeigt in seinem Beitrag auf, dass uns Falsches Erkenntnis liefern kann. Er veranschaulicht, wie Goodmans Symboltheorie zu erklären vermag, weshalb ein Erkenntnisgewinn durch Kunstfälschung möglich ist. Im Anschluss daran berührt Herrmann in seinem Beitrag die These, dass auch Wahres Erkenntnis erweitern kann. Verstehen wir Wahrheit als etwas, das induktiv nicht erschlossen werden kann, ist dies gar nicht so trivial. Ausgehend von James' pragmatistischem Wahrheitsverständnis argumentiert Herrmann dafür, dass wir besser bedient sind, wenn wir der Wahrheit einen relativen Wert beimessen.

Analog dazu wird in der praktischen Philosophie die Frage aufgeworfen, ob nicht auch moralische Werte relativ sein könnten. Im zweiten Teil *Normen des Sollens* nimmt sich David Lussi eben diesem Thema an. In seinem Essay erklärt er, weshalb wir einander Ratschläge geben, auch wenn das, was wir tun sollten, von unseren eigenen epistemischen Perspektiven abhängt. Ferner argumentiert Timo Junger in seinem Beitrag gegen die These, dass notwendigerweise zu viel gefordert wird, wenn eine Moraltheorie supererogatorisches Handeln verlangt. Auf Grundlage verschiedenen Begriffsverständnisse bringt er damit wertvolle Ordnung in das bunte Sammelsurium diverser Ansätze.

Neben konzeptueller Ordnung sind auch konkrete Normen wertvoll. Unter dem Titel *Wertvolle Normen* befassen sich Dela Wälti und Sarah Heinzmann mit eben solchen. Zum einen begründet Wälti in ihrem Beitrag, dass wir künftige Generationen berücksichtigen müssen, um gerechte Rechtsnormen zu schaffen. Zum andern zeigt Heinzmann in ihrem Essay auf, dass der Rechtsbegriff von Genozid überdacht werden muss, weil auch politische Gruppen unter den Schutz dieser Norm fallen sollten.

Zu welchen Gruppen wir gehören, ist nicht immer leicht auszumachen, befinden wir uns doch häufig zwischen verschiedenen Normen. Im abschliessenden Teil dieses Hefts sind unter dem Titel *Normativität und Identität* das Fachwort und ein weiterer Beitrag vereint. Im Fachwort legt Rebekka Hufendiek nach den Regeln der analytischen Kunst dar, inwiefern Emotionen in naturalistischen Ansätzen normativ verstanden werden können und welche Rolle dabei die Einbettung eines Individuums in seiner Umgebung spielt. Auch in Wehrlins Beitrag spielt die Individualität und deren Einbettung eine zentrale Rolle. Während die Autorin der Frage nachgeht, wie der Begriff der Frau definiert werden muss, damit trans-Frauen darunterfallen, zeigt sie auf, wie Begriffsüberarbeitungen zur Harmonisierung von Individualität und Normativität beitragen können.

Nun wünsche ich Ihnen eine wertvolle Lektüre und hoffe, Sie erhaschen einen Einblick in das Schaffen von Studierenden und Dozierenden unseres Instituts, welches zwar der philosophischen, womöglich aber nicht immer ganz der alltäglichen Norm entspricht.

Herzlichst, Vera Moser
im Namen des Redaktionsteams

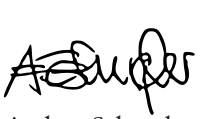

Timo Junger


Isabel Käser


Vera Moser


Aleksandar Nikolić


Xenia Schmidli


Andrea Schnyder

„Wie die zahlreichste Bibliothek, wenn ungeordnet, nicht so viel Nutzen schafft, als eine sehr mäßige, aber wohlgeordnete; eben so ist die größte Menge von Kenntnissen, wenn nicht eigenes Denken sie durchgearbeitet hat, viel weniger werth, als eine weit geringere, die aber vielfältig durchdacht worden."

Arthur Schopenhauer
Selbstdenken, 1851

"We have in fact two kinds of morality side by side: one which we preach, but do not practice, and another which we practice, but seldom preach."

Bertrand Russel
On Ethics, Sex, and Marriage, 1987

Inhaltsverzeichnis

Wert und Erkenntnis

Normen des Sollens

Wertvolle Normen

Normativität und Identität

Der Wert der Fälschung

Eine Anwendung der Symboltheorie Nelson Goodmans

1. Einleitung

Wolfgang Beltracchi wurde im Oktober 2011 wegen Bandenbetrugs zu sechs Jahren Haft verurteilt.[1] Der 1953 geborene, deutsche Maler täuschte während rund 40 Jahren den internationalen Kunstmarkt, indem er nach eigener Angabe rund dreihundert Bilder malte, mit falscher Signatur versah und als vermeintliche Originale verschiedener Künstler*innen an zahlreiche Galerien verkaufte. Nachdem Beltracchis Täuschung aufgeflogen war, verloren die vermeintlichen Originale natürlich über Nacht an finanziellem Wert. Die getäuschten Gläubiger*innen wurden für den finanziellen Schaden entschädigt. Viele Käufer*innen wollten die Bilder dennoch nicht behalten. Das wirft unter anderem die Frage auf, woran sich der Wert einer Fälschung[2] messen lässt. Zweifelsohne gelten Fälschungen für viele Menschen und für den Kunstmarkt speziell als wertlos. In *Die Kunst der Fälschung*, einem Dokumentationsfilm zum Fall Beltracchi, stützt James Roundell als Vertreter der Society of London Art Dealers diese Behauptung, indem er anmerkt, dass Beltracchi einerseits nichts Neues und andererseits gewiss nichts geschaffen habe, was von Wert wäre. Roundells Urteil bezieht sich auf den künstlerischen Wert, der den finanziellen Wert mitbestimmt, da im Kontext des internationalen Kunstmarkts nur wertvoll ist, was authentisch ist. Hinreichende und notwendige Bedingung dafür, ein authentisches Max Ernst Bild zu sein, ist die Eigenschaft zu besitzen, von Max Ernst gemalt worden zu sein. „La Forêt II", ein Bild von Beltracchi, hat diese Eigenschaft nicht. Dennoch wurde „La Forêt II" auf dem internationalen Kunstmarkt als authentisches Max Ernst Bild verkauft. Werner Spies, der damals führende Max-Ernst-Experte, hielt „La Forêt II" für ein von Max Ernst gemaltes Bild. Ein Bild kann also als authentisch gelten, obwohl es kein Original ist.

Durch eine symboltheoretische Betrachtung lässt sich erklären, welche Symboleigenschaften ein Bild haben muss, damit es als authentisch gelten und erfolgreich täuschen kann. Symboleigenschaften sind die Eigenschaften eines Bildes, auf die das Bild in einer bestimmten Situation referiert. Der US-amerikanische Philosoph Nelson Goodman hat mit *Languages Of Art* eine umfassende Symboltheorie vorgelegt. Der Begriff des Symbols ist darin von Goodman so definiert, dass eine allgemeine Anwendung möglich ist. Wir können also auch Gegenstände wie Bilder als Symbole verwenden. Durch die Verwendung der Bilder als Symbole wird in mehrfacher Hinsicht deutlich, dass die besagten Bilder von Beltracchi nicht wertlos sind. Fälschungen dieser Art sind auf eine einzigartige Weise epistemisch wertvoll. Der Erkenntnisgewinn, den ich dabei im Blick habe, ist nur möglich, gerade weil es Fälschungen dieser Art sind. Bevor wir die Bilder als Symbole verwenden, erkläre ich zunächst, um welche Art von Bildern es mir dabei geht.

2. Arten der Fälschung

Fälschungen lassen sich grob in Kopien und Ergänzungsfälschungen unterteilen. Erstere können erstellt werden, indem ein Original möglichst genau nachgemalt wird. Kopien werden oft legal hergestellt und verkauft. Sie sind erschwingliche Alternativen zu den Originalen. Die Herstellung einer Kopie ist zwar eine gute Übung, um die eigenen Malfertigkeiten zu schulen, für eine Fälscher*in, die Kopien als vermeintliche Originale verkaufen will, aber denkbar ungeeignet. In den meisten Fällen ist genau dokumentiert, wem ein bestimmtes Bild gehört

1 Alle biografischen Inhalte zu Beltracchi sind aus *Selbstporträt* oder dem Dokumentarfilm *Die Kunst der Fälschung*.

2 Gemeint sind vor allem Fälschungen von Kunstwerken. Es scheint so, dass andere Fälschungen, gerade weil sie Fälschungen sind, einem Individuum innerhalb eines bestimmten Umfeldes beispielsweise Prestige verleihen. So kann eine gefälschte Gucci-Tasche in Form von sozialem Kapital wertvoll sein.

und wo es sich gerade befindet. Die Chance, mit einer Kopie zu täuschen, ist deshalb praktisch ausgeschlossen.

Für Fälscher*innen sind die Ergänzungsfälschungen, für die sich auch Beltracchi entschieden hat, wesentlich erfolgsversprechender. Beltracchi kopierte keine Originale. Er malte neue Bilder und ergänzte damit Werklücken. Es gibt zwei Arten der Ergänzungsfälschungen. Die erste Art ergänzt insofern, als dass sie verschollene Bilder ersetzt. Gemeint sind Bilder, deren Existenz belegt ist, aber niemand darüber Bescheid weiss, wie das Bild genau aussieht. Zum Beispiel deshalb nicht, weil eine bestimmte Künstlerin dieses Bild bloss mit einer kurzen Beschreibung in einem Briefwechsel erwähnt hat, sonst aber alle Anhaltspunkte bezüglich des Aufenthaltsortes des Bildes im Laufe der Zeit verloren gegangen sind. Die zweite Art ergänzt ein Werk mit komplett erfundenen Bildern. Angenommen, von einer Künstlerin ist von 1912 bis 1915 die Gelb-Phase bekannt und ab 1917 die Rot-Phase. Dann hätte Beltracchi möglicherweise das Werk in den zwei Jahren dazwischen mit der Orange-Phase ergänzen können.

zungsfälschung keine Vorlage. Ergänzungsfälschungen bedingen deshalb neben dem technischen Handwerk, was auch beim Kopieren verwendet wird, eine bestimmte Art des kreativen Schaffens. Sie müssen das bekannte Werk einer Künstler*in plausibel ergänzen. Allein deshalb ist die Behauptung, Ergänzungsfälscher*innen hätten nichts Neues, nichts von künstlerischem Wert geschaffen, unbegründet. Damit auch der epistemische Wert dieser Fälschungen ersichtlich wird, müssen wir zuerst Klarheit über den Begriff der Handschrift schaffen. Ich verwende „Handschrift“ als Arbeitsbegriff und bin mir bewusst, dass damit einige Unklarheiten einhergehen können – aber dazu später mehr.

3. Handschrift, Stil und Charakter

Das Bild „La Forêt II“, das angeblich 1927 von Max Ernst gemalt wurde und zu einer Reihe von *Wäldern*[3] aus diesem Jahr gehört, galt lange als verschollen, bis Beltracchi es 1998 gemalt hatte. Bisher sind insgesamt sieben Bilder bekannt, die als authentische Max Ernst Bilder galten, aber von Beltracchi gemalt wurden. Werner Spies beur-

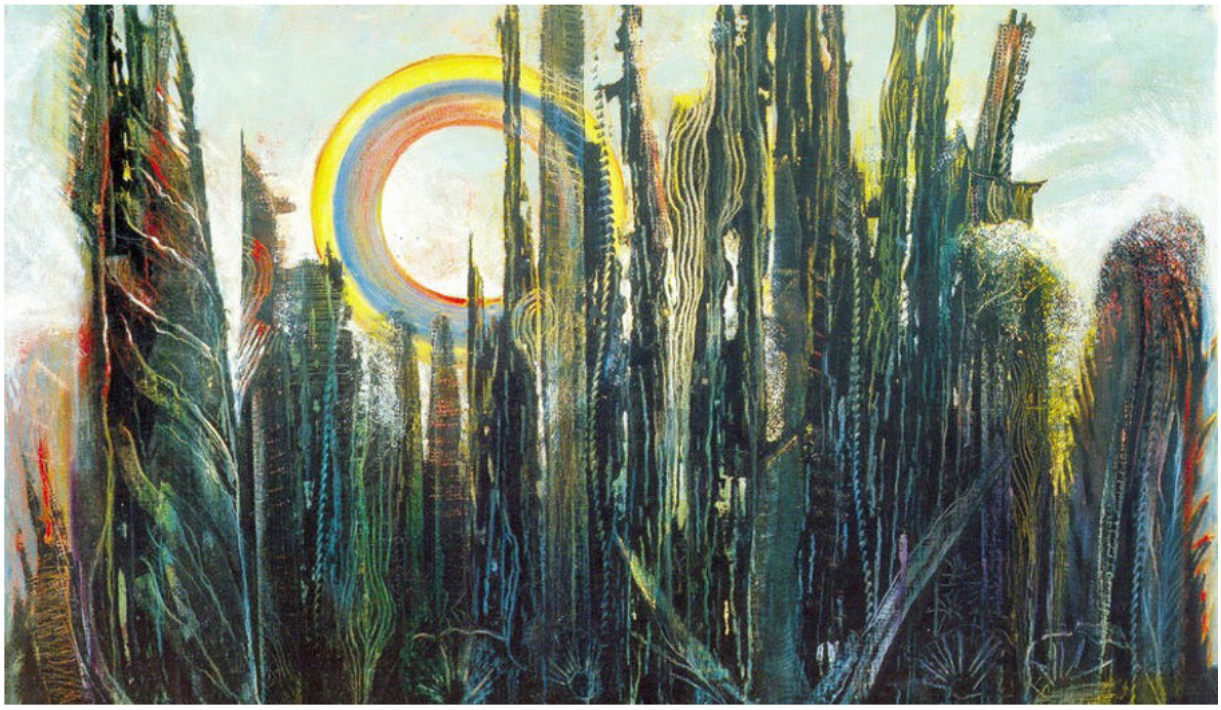


Abb.1 – „La Forêt II“ 1998 von Wolfgang Beltracchi

Beltracchi hat Ergänzungsfälschungen beider Arten geschaffen. Wie die Kopie ist auch dieErgänzungsfälschung ein Original der Person, die sie gemalt hat, und wird erst zur Fälschung, wenn vorgegeben wird, dass sie von einer anderen Person gemalt wurde. Beltracchi hat dazu die entsprechenden Unterschriften nachgeahmt, weshalb er unter anderem wegen Urkundenfälschung verurteilt wurde. Anders als bei der Kopie gibt es für eine Ergänteilte sie alle als authentisch. Gemäss Beltracchi seien seine Bilder von den Expert*innen unter anderem deshalb für authentisch gehalten worden, weil sie die richtige Handschrift aufweisen. Er habe nicht Bilder im Stil von Max Ernst gemalt, sondern in dessen Handschrift. In Beltracchis Redeweise bezeichnet Stil eine bestimm-

3 Die kursive Schreibweise soll deutlich machen, dass es sich dabei um Bilder handelt.

te Epoche. Demnach gibt es als Stil beispielsweise den Expressionismus oder die Renaissance. Über die Unterscheidung zwischen Stil und Handschrift lässt sich streiten. Im Folgenden werde ich sie beibehalten, weil auch andere Stilbegriffe nicht genau das einfangen, was ich mit der Handschrift meine.

So verwendet Nelson Goodman einen Stilbegriff, demzufolge der Stil eines Kunstwerkes[4] von jenen Merkmalen abhängt, die für Autor*in, Zeit, Ort oder Schule charakteristisch sind (Goodman 1978, 38). Goodmans Stilbegriff schliesst demnach Autor*innenschaft mit ein. Autor*innenschaft ist ebenfalls zentral für eine bestimmte Handschrift, die ein Bild symbolisiert. Die Handschrift einer bestimmten Person wird nur durch ebendiese Person geprägt. Goodmans Stilbegriff ist in einem gewissen Sinne weiter gefasst als mein Begriff der Handschrift, weil nach Goodman Stilunterschiede auch durch verschiedene Autor*innen bewirkt werden können. Er nennt als ein Beispiel die Unterschiede zwischen Barock und Rokoko (Goodman 1978, 34).

Ein Teil dessen, was ich mit Handschrift meine, sind charakteristische Merkmale der individuellen Malweise einer Künstler*in. Diese lassen sich durch das Beantworten von Fragen folgender Art erfassen: Wurde mit der rechten Hand gemalt? Wie schnell wurde das Bild gemalt? Wie dick oder stark wurde die Farbe aufgetragen? Zur individuellen Malweise gehören auch die Komposition und die verwendeten Techniken. Neben der Malweise sind in meinem Verständnis die verwendeten Materialien ebenfalls Teil der Handschrift. Dies ist ein weiterer Grund dafür, der Begriff der Handschrift von Goodmans Stilbegriff zu unterscheiden. Für Goodman sind Eigenschaften des Rahmens oder der Pigmente keine stilistischen Merkmale (Goodman 1978, 34–35). Ich zähle sie dennoch zur Handschrift. Wenn wir annehmen, dass ein Bild die korrekte Handschrift aufweisen muss, um als authentischer *Wald* von Max Ernst aus dem Jahre 1927 zu gelten, dann muss auch der Rahmen und die Leinwand des Bildes zu der entsprechenden Zeit passen, ansonsten wird die Täuschung kaum gelingen. Die aufgezählten charakteristischen Merkmale bilden eine Teilmenge der Symboleigenschaften des Bildes. Dass ein Bild die korrekte Handschrift aufweist, heisst hier nichts anderes, als dass das Bild die Handschrift Max Ernsts exemplifiziert. Die Exemplifikation als Art der Referenz erkläre ich im Unterkapitel 4.2.

Auf eine andere Weise ist Goodmans Stilbegriff eher eng gefasst, da er fordert, dass ein Kunstwerk nur dann dem Stil einer bestimmten Autor*in, einer Region und einer Zeit angehören kann, wenn das Kunstwerk von dieser Autor*in, in der entsprechenden Region und der entsprechenden Zeit geschaffen wurde (Baumberger und Brun 2013, 153). Nur ein Kunstwerk aus dem Expressionismus kann einen expressionistischen Stil haben. Bei Christoph Baumberger und Georg Brun findet sich ein Charakterbegriff, der Goodmans Stilbegriff ergänzt, indem sie den Stil nach Goodman als einen besonderen Charakter eines bestimmten Bauwerks ansehen. Baumberger und Brun definieren „Charakter“, um auf eine bestimmte Weise über die Identität und Teilidentitäten von Bauwerken zu sprechen. Der Begriff des Charakters lässt sich auch auf Gemälde oder Bilder anwenden. Frei nach der Definition Baumbergers und Bruns ist jede nicht-leere Teilmenge der Symboleigenschaften eines Bildes, ein Charakter dieses Bildes (Baumberger und Brun 2013, 151). Mit dieser Definition lässt sich präziser über Bauwerke oder Bilder sprechen. Ein bestimmtes Bild hat zwar keinen expressionistischen Stil, weil es nicht aus dem Expressionismus stammt. Wir können dennoch sagen, dass es einen expressionistischen Charakter hat, weil es einen neoexpressionistischen Stil hat (Baumberger und Brun 2013, 151–154). Wie den Stil verstehe ich auch die Handschrift als eine Teilmenge der Symboleigenschaften eines Bildes und damit als einen bestimmten Charakter dieses Bildes. Entsprechend dem genannten Beispiel können wir sagen, dass ein Bild einen surrealistischen Charakter hat, weil es die Handschrift Max Ernsts exemplifiziert.

4. Zwei Relationen der Bezugnahme

Damit ein Bild eine Handschrift exemplifiziert, müssen wir das Bild als Symbol verwenden. Damit klar wird, was damit gemeint ist, müssen wir erst die Variationen der Referenz verstehen, die Goodman in seiner Theorie erläutert. Referenz kommt durch eine Relation der Bezugnahme zwischen einem Symbol und einem Objekt, auf welches referiert wird, zustande. Goodman unterscheidet unter anderem zwischen zwei verschiedenen Relationen. Mittels Denotation lassen sich Gegenstände im weitesten Sinne bezeichnen. Daneben beschreibt Goodman die Bedingungen für die Bezugnahme mittels Exemplifikation. Um zu zeigen, inwiefern Fälschungen epistemisch wertvoll sind, werde ich später immer wieder die exemplifizierende Relation erwähnen. Da jedoch Denotation eine notwendige Bedingung für die Exemplifikation ist, bietet es sich an, die Erläuterung mit der Denotation anzufangen.

4 Ich schreibe „Kunstwerk“ anstelle von „Werk“, weil ich ein Einzelding meine. Die Definition gilt aber auch für Werke, die nicht eindeutig als Kunstwerke angesehen werden wie beispielsweise manche Bauwerke. Mit „Werk“ bezeichne ich weiterhin ein Gesamtwerk einer Künstler*in.

4.1 Denotation

Wir verwenden Denotation dann, wenn wir mit einem Symbol *a* auf einen Gegenstand *b* referieren, *a* für *b* steht oder *a* ein *b* repräsentiert. Wenn ein Symbol einen Gegenstand denotiert, dann referiert es auf diesen Gegenstand. Ein Gegenstand, der denotiert wird, bezeichnet Goodman als Denotat. Als Symbol und als Gegenstand kommen nahezu alle Entitäten in Frage. Eine Ähnlichkeitsbeziehung zwischen dem Symbol und dem Gegenstand, auf den referiert wird, ist laut Goodman weder eine hinreichende noch eine notwendige Bedingung dafür, dass eine Referenzbeziehung hergestellt werden kann (Goodman 1976, 5). So kann die Buchstabenfolge „Fuchs" für einen beliebigen Fuchs stehen und es gelingt mit „Fuchs" auf einen Fuchs zu referieren, obwohl der Fuchs der Buchstabenfolge nicht im Geringsten ähnelt. Die Verwendung eines Symbols entsprechend der Konvention innerhalb eines etablierten Symbolsystems ist allein relevant dafür, ob es gelingt, mit einem Symbol *a* auf einen Gegenstand *b* zu referieren. Im Symbolsystem der deutschen Sprache kann mit der Buchstabenfolge „Ball" sowohl auf ein rundes Sportgerät wie auch auf einen Tanzabend referiert werden. Hingegen wird es mit „Klorfz" momentan nicht gelingen, verständlich auf ein Denotat zu referieren. Der jeweilige Kontext, in dem *a* als Symbol verwendet wird, ist insofern entscheidend dafür, ob es gelingt, die gewünschte Referenzbeziehung zu *b* herzustellen, als dass durch den Kontext bestimmt ist, in welchem Symbolsystem *a* verwendet wird und welche Konventionen dafür gelten. In einem bestimmten Kontext kommt nur ein Symbolsystem zur Anwendung, aber ein bestimmtes Symbolsystem kann in mehreren Kontexten verwendet werden.

Laut Goodman gibt es auch Symbole, die nichts denotieren. Für „Einhorn" lässt sich kein Denotat finden, da es keine Einhörner gibt. Trotzdem haben wir ein Bild davon, wie ein Einhorn aussehen muss, um als Einhorn und nicht beispielsweise als Kobold zu gelten. Was wir mit „Einhorn" meinen, lernen wir mit Hilfe von Einhorn-Bildern. Ein solches Einhorn-Bild zu sein, heisst im Sinne Goodmans, von dem Etikett (*label*) „Einhorn-Bild" denotiert zu werden, anstatt selbst etwas zu denotieren. Wir unterscheiden ein Einhorn-Bild von einem Kobold-Bild genauso wie wir Tische von Stühlen unterscheiden, die ebenfalls nichts denotieren (Goodman 1984, 60). Die Referenzrichtung verweist bei Denotation vom Symbol aus auf das Denotat. Verwenden wir ein Etikett als Symbol, verweist das Etikett auf das Denotat. Ein Etikett kann vereinfacht als Prädikat im weitetesten Sinne verstanden werden, obwohl Goodman erwähnt, dass Prädikate lediglich Etiketten seien, die zu wortsprachlichen Symbolsystemen (*linguistic systems*) gehören, in denen Etiketten üblicherweise Gegenstände oder Eigenschaften bezeichnen (Goodman 1976, 57). Bei der Exemplifikation läuft die Referenzrelation in die andere Richtung. Das heisst, Exemplifikation kommt durch die Verwendung eines Musters als Symbol für ein entsprechendes Etikett zustande. Das Muster exemplifiziert dabei das Etikett.

4.2 Exemplifikation

Exemplifikation lässt sich gut anhand eines ausgestanzten Musters einer bestimmten Ledersorte erklären, wie sie oft von Bandagist*innen oder Schuhmacher*innen verwendet werden. Das kreisförmige Muster dient dazu, gewisse Eigenschaften der Ledersorte zu exemplifizieren. So zum Beispiel die Reissfestigkeit, die Farbe und die Haptik der Ledersorte. Nicht aber, dass die Haut an einem Montag von Katarina gegerbt wurde, obwohl das ebenfalls eine Eigenschaft des Leders sein könnte. Ein ungefärbtes Ziegenleder ist beispielsweise beige und hat eine glatte Narbenseite. Ein Muster dieses Ziegenleders muss also die Etiketten „beige" und „glatte Narbenseite" exemplifizieren. Dazu muss es selbst beige sein und eine glatte Narbenseite haben. Ein Muster kann nur Etiketten exemplifizieren, die es selbst besitzt und auf die es gleichzeitig referiert (Goodman 1976, 53). „Besitzen" heisst in Goodmans Verständnis, dass das Muster die Eigenschaft haben muss, die das Etikett beschreibt, welches exemplifiziert werden soll. So kann das Ledermuster auch verwendet werden, um das Etikett „kreisförmig" zu exemplifizieren oder um zu exemplifizieren, was mit einem Locheisen gemacht werden kann, wenn das Ledermuster mit einem Locheisen ausgestanzt wurde. Welche Etiketten das Muster als Symbol exemplifiziert, ist abhängig vom jeweiligen Kontext und dem Symbolsystem, in welchem das Muster verwendet wird. Dabei spielt es keine Rolle, ob das Muster die Eigenschaft wörtlich oder metaphorisch besitzt.

Um gewisse Etiketten zu exemplifizieren, muss das Muster gleichzeitig von diesen Etiketten denotiert werden. Denotation ist demnach eine notwendige Bedingung der Exemplifikation. Während Denotation eine einfache Referenzbeziehung ist, die vom Symbol aus auf das Denotat verweist, bestehen bei der Exemplifikation zwei Beziehungen. Eine vom Muster zum Etikett, das exemplifiziert wird, und eine zweite vom Etikett zum denotierten Muster (Goodman 1976, 59). Während nahezu alles denotiert werden kann, sind es ausschliesslich Etiketten, die exemplifiziert werden können. Damit *a* als ein Muster für *b* verwendet werden kann, müssen folgende Bedingungen der Exemplifikation erfüllt sein: *a* exemplifiziert *b*, wenn (1) *a* von *b* wörtlich oder metaphorisch denotiert wird, (2) *a* wörtlich oder metaphorisch *b* besitzt und (3) *a* auf *b* referiert (Goodman 1976, 95).

5. Erkenntnis durch Exemplifikation

Abb.2 – „La grande Forêt“ 1927 von Max Ernst

Abb.3 – „Forêt“ 1927 von Max Ernst

Goodmans Symboltheorie lässt sich nun problemlos auf Gemälde anwenden. Wir können sagen, dass ein Gemälde von unendlich vielen Etiketten denotiert wird. Die meisten dieser Etiketten denotieren keine Merkmale, die zur Beurteilung der Authentizität wichtig sind. So sind zum Beispiel die Etiketten „wurde von zwei Personen aufgehängt“ oder „Landeplatz für eine Fliege am 28.03.1978“ dafür irrelevant. Bestimmte charakteristische Merkmale der *Wälder* von Max Ernst können aber als Symbole verwendet werden, um die Handschrift Max Ernsts[5] zu exemplifizieren, und sind damit für die Beurteilung der Authentizität seiner Bilder relevant. Eine symboltheoretische Erklärung auf die eingangs gestellte Frage nach den Eigenschaften, die ein Bild haben muss, um als authentisch gelten zu können, wäre zu sagen, dass ein Bild die Handschrift Max Ernsts exemplifizieren muss, um als authentisches Max Ernst Bild gelten zu können.

5.1 Bilder als Muster

Betrachten wir zuerst ausschliesslich die von Max Ernst gemalten *Wälder*, lassen sich einfach charakteristische Merkmale finden, die, wie erwähnt, zur Handschrift gehören könnten. Auf den Abbildungen 2–4 lässt sich die typische Komposition leicht erkennen, die sich grob in Vorder- und Hintergrund einteilen lässt. Der jeweilige Vordergrund der einzelnen Bilder wird stets von einem komplexen, gestrüppartigen Element in eher dunklen Farbtönen dominiert. Max Ernst hat die Hintergründe im Vergleich zu den Vordergründen eher ruhig gestaltet. Weiter ist sämtlichen *Wäldern* ein kreisförmiges Element gemeinsam, das das Etikett „monochrome Sonne“ exemplifiziert. Beim Malen seiner *Wälder* hat Max Ernst die Techniken der Frottage[6] und Grattage[7] verwendet, für die er bekannt war. Teile seiner *Wälder* können als Muster verwendet werden, welche die Etiketten „Frottage“ und „Grattage“ exemplifizieren. Auch für die gewählten Farben lassen sich entsprechende Etiketten finden, die von den *Wäldern* exemplifiziert werden und diese oder Teile davon denotieren. Ich habe hier nur einige mögliche Etiketten der Handschrift Ernsts ausformuliert.

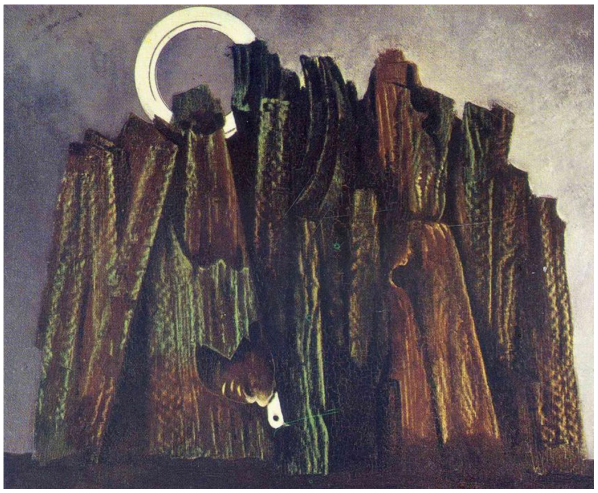


Abb.4 – „Forêt sombre et oiseau“ 1927 von Max Ernst

Als Experte für Max Ernst waren Werner Spies die charakteristischen Merkmale eines *Waldes* von Ernst natürlich bekannt. Er kannte dessen Handschrift, was seine Einschätzung der Bilder, die er vorgelegt bekam, mehr oder weniger bewusst beeinflusst hatte. Die *Wälder* von Beltracchi mussten entsprechend Max Ernsts Handschrift exemplifizieren, um Werner Spies zu täuschen.

5 Im Folgenden ist nur die Handschrift in Ernsts *Wäldern* gemeint. Sein restliches Werk wird grösstenteils von Etiketten denotiert, die nicht von seinen *Wäldern* oder einzelnen Merkmalen davon exemplifiziert werden können.

6 Frottage: Ein Blatt oder eine Leinwand wird auf einen unebenen Untergrund gelegt. Durch Abreiben von Farbe wird die Struktur des Untergrunds sichtbar.

7 Grattage: Es werden in die noch nassen Farbschichten Strukturen eingekratzt.

Welche Etiketten genau für Spies entscheidend waren, können wir nicht abschliessend bestimmen. Jedoch wird es nicht gereicht haben, dass ein *Wald* bloss „Rahmen aus 1927“ exemplifiziert hat. Ebenso unwahrscheinlich scheint es, dass Werner Spies die Authentizität von „La Forêt II“ an einem einzigen Merkmal festmachte.[8] Die Echtheit eines Ziegenlederes bestimmt sich auch nicht bloss anhand eines einzigen Merkmals. Wir können die Merkmale von echtem Ziegenleder mit einem Muster lernen, das die dafür relevanten Etiketten exemplifiziert. Unter Umständen können wir Merkmale eines Ziegenlederes auch anhand eines Musters lernen, das selbst kein Ziegenleder ist. Beispielsweise können wir die schraffierte Oberfläche eines Saffianleders anhand eines veganen Musters lernen.

Genauso können wir einen *Wald* von Beltracchi als Muster heranziehen, um Erkenntnisse über die *Wälder* von Ernst zu gewinnen. Es scheint nicht plausibel, zu sagen, dass wir nur anhand von *Wäldern*, die von „von Max Ernst gemaltes Bild“ denotiert werden, lernen können, wie ein *Wald* von Max Ernst aussieht. Die Annahme, dass wir nur von authentischen Bildern Erkenntnisse über die *Wälder* von Max Ernst gewinnen können, ist falsch, da Erkenntnisgewinn auch anhand nicht-authentischer Bilder gelingt. Zum Beispiel können wir auf einem Foto eines Gemäldes erkennen, wie einzelne Elemente auf dem darauf abgebildeten Gemälde angeordnet sind. Gewisse Merkmale nicht-authentischer Bilder tragen dazu bei, dass sie für authentisch gehalten werden, was die Einschätzung des Experten im Fall Beltracchi beweist. Goodman schreibt, dass wir Merkmale von Bildern finden können, anhand derer es uns gelingen kann, die entsprechenden Bilder dem Werk einer bestimmten Person zuzuschreiben und sie von den Bildern, die nicht zu ebendiesem Werk gehören, zu unterscheiden (Goodman 1976, 109).

5.2 Eine Handschrift erkennen

Wenn wir lernen, *Beltracchis* von *Nicht-Beltracchis* zu unterscheiden, teilen wir dabei den Mustern, also den einzelnen Bildern, entweder das Etikett „Beltracchi“ oder „nicht-Beltracchi“ zu. Laut Goodman ist die Anzahl Muster, die uns zur Verfügung steht, entscheidend dafür, wie zuverlässig uns diese Unterscheidung gelingt (Goodman 1976, 109). Goodman führt, um dies zu zeigen, als Beispiele die gefälschten *Vermeers* an, die Van Meegeren in den 1930er Jahren gemalt hat. Für eine Expert*in, die zu Beginn nur ein Bild von Van Meegeren vor sich hatte, war es verhältnismässig schwierig, zu beurteilen, ob dieses Bild genug nach einem *Vermeer* aussah, um der bereits bekannten Menge der *Vermeers* zugeteilt werden zu können. Das heisst letztlich, dass es schwierig war, zu beurteilen, ob das Bild die Handschrift Vermeers exemplifizierte. Je mehr Bilder Van Meegeren erfolgreich als *Vermeers* verkaufte, desto einfacher war es, mit weiteren Bildern zu täuschen, da er mit jedem Bild, das es in die Menge der *Vermeers* schaffte, die Handschrift zu seinem Vorteil veränderte. Das gilt nur, solange die Täuschung noch nicht aufgeflogen ist. Ansonsten wäre es tatsächlich so, dass die Handschrift auch von mehreren Autor*innen verändert werden kann. Mittlerweile wurden die *Van Meegerens* aus der Menge der *Vermeers* entfernt und bilden eine eigene Menge. Eine zuverlässige Beurteilung der Bilder wird dadurch wesentlich leichter, da die Unterschiede sofort auffallen (Goodman 1976, 111). Es scheint unvorstellbar, dass eine Person, die beide Werke kennt, einen *Van Meegeren* für einen *Vermeer* halten kann. Dies, weil die *Vermeers* von Van Meegeren die Handschrift Vermeers nicht exemplifizieren.

Abb.5 – „Der Astronom“ 1668 von Vermeer

Ein Vergleich zwischen zwei Bildern macht einige Unterscheide deutlich. Van Meegeren hat flächiger und gröber gemalt. Die Gesichter und Hände wirken zu klobig, um von Vermeer gemalt zu sein. Das heisst aber nicht, dass die *van Meegerens* keine Etiketten der Handschrift Vermeers exemplifizieren. Wir können

8 Ich meine damit, dass die Handschrift Max Ernsts nicht bloss aus einem einzigen Merkmal besteht. Dass ein Bild die Handschrift Max Ernsts exemplifiziert, kann natürlich auch als ein einziges Merkmal verstanden werden, das womöglich für die Beurteilung der Authentizität reichen würde. Das ist aber nicht das, was ich hier meine.

sagen, dass die *van Meegerens* dennoch einzelne charakteristische Merkmale der Handschrift Vermeers exemplifizieren. Bei den *van Meegerens* reduziert sich die Erkenntnis bezüglich Vermeers Werk möglicherweise darauf, dass Vermeer die Lichtquelle oft linksoben platziert hat und dass er oft Menschen gemalt hat. Die symboltheoretische Analyse erlaubt uns ein Stück weit, Kriterien für eine gute Fälschung anzugeben. Je mehr Etiketten der Handschrift exemplifiziert werden, desto besser ist die Fälschung. Anhand von guten Fälschungen können wir lernen zu erkennen, welches die charakteristischen Merkmale der Handschrift einer bestimmten Künstler*in sind. Wir können anhand von Beltracchis „La Forêt II" lernen, wie ein *Wald* von Ernst aussieht und „La Forêt II" als Muster verwenden, um andere Bilder als *Wälder* von Ernst zu identifizieren. Dies ist die erste Art der Erkenntnis, die wir durch Fälschungen gewinnen können. Sie gilt sowohl für Kopien wie auch für Ergänzungsfälschungen. Guten Ergänzungsfälschungen kann jedoch noch auf eine andere Weise epistemischer Wert zugeschrieben werden, nachdem sie als Fälschungen entlarvt wurden.

Abb.6 – Ausschnitt aus „Das Letzte Abendmahl" 1939 von Van Meegeren

5.3 Ein Werk ergänzen

Damit ein *Wald* von Beltracchi als Ergänzungsfälschung täuschen konnte, war es notwendig, dass sich dieser in die Reihe der *Wälder* von Max Ernst einfügte. Er musste das Werk von Ernst auf eine plausible Weise ergänzen. Die Handschrift Max Ernsts gab also gewissermassen den Rahmen für die *Wälder* von Beltracchi vor. Das ist ein entscheidender Unterscheid zur Arbeitsweise von Max Ernst, dem als Künstler keine Grenzen gesetzt waren und der sich mit jedem Bild neu erfinden konnte. Wie bereits erwähnt, hat Beltracchi tatsächlich viele Elemente der *Wälder* von Max Ernst übernommen. Aber seine Ergänzungsfälschungen weisen zusätzlich auch neue Elemente auf, wodurch sie möglicherweise besonders authentisch wirkten. Gleichzeitig wird eine Ergänzungsfälschung durch das Exemplifizieren von neuen Etiketten epistemisch besonders wertvoll. Vergleichen wir „La Forêt II" mit den *Wäldern* von Max Ernst, fallen sofort klare Unterschiede auf. Die monochromen Sonnen von Max Ernst hat Beltracchi beispielsweise durch eine mehrfarbige Sonne ersetzt. Zudem ist „La Forêt II" insgesamt farbenfroher gestaltet und wirkt heiterer. Die Ergänzungsfälschungen von Beltracchi kontrastieren das Werk von Ernst auf eine mögliche Art. Indem Beltracchi einen bunten *Wald* malt, zeigt er, wie ein *Max Ernst* auch noch hätte aussehen können. Einigen wird vielleicht erst durch diesen bunten *Wald* bewusst, wie düster Max Ernst seine Waldreihe gemalt hat. Neben neuen rein ästhetischen Erkenntnissen können wir durch die Kontrastierung zusätzlich etwas über die Bedeutung einzelner Merkmale der Bilder von Max Ernst erfahren. Gleichzeitig treten hier auch die angesprochenen Unklarheiten mit dem Begriff der Handschrift auf.

Ich habe oben angedeutet, dass die Handschrift immer nur durch eine einzige Person verändert wird. Weiter soll ein Gemälde, um als authentisches Bild gelten zu können, die richtige Handschrift exemplifizieren. Wenn wir annehmen, dass „monochrome Sonne" zur Handschrift Ernsts gehört, stellt sich die Frage, wie „La Forêt II" als authentisch gelten konnte. Da „La Forêt II" das Etikett „monochrome Sonne" nicht exemplifiziert, heisst das möglicherweise, dass die Farbe der Sonne für die Beurteilung der Authentizität egal ist. Es ist unplausibel, dass Beltracchi mit seinen Bildern verändert, was wir bisher als die Handschrift Max Ernsts verstanden haben. Das würde bedeuten, dass „monochrome Sonne" gar nie als Teil der Handschrift exemplifiziert wurde. Oder aber, dass ein Bild, um als authentisch zu gelten, nicht alle Etiketten der Handschrift, sondern nur eine Teilmenge davon exemplifizieren muss. Wir können einsehen, dass für die Beurteilung der Authentizität nicht alle charakteristischen Merkmale einer Handschrift gleich wichtig sind. Angewandt auf die Handschrift Max Ernsts und die monochrome Sonne, würde dann gelten, dass jene Merkmale seiner *Wälder*, die die Anordnung der Elemente betreffen, für das Erkennen eben dieser Handschrift wichtiger sind als farbliche Merkmale. Welche Merkmale Teil der Handschrift sind, lässt sich aber weiterhin nicht abschliessend sagen.

6. Schlusswort

Das Ziel dieses Beitrages war es aufzuzeigen, dass Fälschungen ein epistemischer Wert zugeschrieben werden kann. Dieser kommt zu Stande, indem wir Bilder als Muster sehen, die eine bestimmte Handschrift exemplifizieren. Damit eine Fälschung als authentisches Original einer Künstler*in gelten kann, muss die Fälschung die Handschrift dieser Künstler*in exemplifizieren. Folglich können wir anhand der Fälschung die charakteristischen Merkmale des Werks oder einer Werkphase einer Künstler*in lernen. Ergänzungsfälschungen können wir einen weiteren Wert zuschreiben. Indem diese ein Werk mit neuen Elementen auf eine plausible Art ergänzen, zeigen sie uns eine mögliche Entwicklung einer Künstler*in. Nachdem die Fälschung als solche entlarvt wurde, kontrastiert sie das bestehende Werk.

Literatur

- Baumberger, Christoph und Georg Brun. 2013. „Identität, Charakter und Stil von Bauwerken." In Baumberger, Christoph (Hrsg.): *Architekturphilosophie: Grundlagentexte*, 141–166. Paderborn: Mentis.
- Beltracchi, Helene und Wolfgang Beltracchi. 2014. *Selbstporträt.* Reinbek bei Hamburg: Rowohlt.
- Goodman, Nelson. 1976. *Languages of Art. An Approach to a Theory of Symbols.* Indianapolis & Cambridge: Hackett.
- ———. 1978. *Ways of Worldmaking.* Indianapolis: Hackett.
- ———. 1984. *Of Mind and Other Matters.* Cambridge, MA & London: Harvard University Press.

Film

- Birkenstock, Arne. 2014. *Die Kunst der Fälschung.* Köln: Fruitmarket Kultur und Medien & Tradewind Pictures.

Bilder

- Abb.1 – Beltracchi, Wolfgang. 1998. *La Forêt II.* Öl auf Leinwand. New York: Privatsammlung. Bild von: http://olivierrossel.ch/less-concentrationmorecollision/collide_more.html
- Abb.2 – Ernst, Max. 1927. *La grande Forêt.* Öl auf Leinwand. 114,5 x 146,5 cm. Basel: Kunstmuseum. Bild von: https://www.pinterest.com/pin/504121752010104358/
- Abb.3 – Ernst, Max. 1927. *Forêt.* Öl auf Leinwand. 114 x 146 cm. Privatsammlung. Bild von: https://www.tagesspiegel.de/kultur/marlene-streeruwitz-ueber-max-ernst-kein-feind-kein-freund-kein-wolf/19285946.html
- Abb.4 – Ernst, Max. 1927. *Forêt sombre et oiseau.* Öl auf Leinwand. 65 x 81 cm. Privatsammlung. Bis 2001 Bloomington: Indiana University Art Museum. Bild von: https://www.pinterest.com/WahooArt-Germany/max-ernst-gem%C3%A4lde/
- Abb.5 – Vermeer, Jan. 1668. *Der Astronom.* Öl auf Leinwand. 51,5 x 45,3 cm. Paris: Louvre. Bild von: https://de.wikipedia.org/wiki/Der_Astronom#/media/Datei:VERMEER_El_astr%C3%B3nomo_(Museo_del_Louvre,_1688).jpg
- Abb.6 – Van Meegeren, Han. 1939. *Das Letzte Abendmahl.* Öl auf Leinwand. Rotterdam: Caldic Collection. Bild von: Edward, Dolnick. 2008. *The Forger's Spell. A True Story of Vermeer, Nazis, and the Greatest Art Hoax of the Twentieth Century.* New York: HarpersCollins Publishers.

Nicola Grossrieder studiert Philosophie in Bern. Sein theoretisches Interesse gilt der Sprachphilosophie, der Philosophie des Geistes und der Metaphysik. Weiter beschäftigen ihn aktuell Fragen bezüglich Verantwortung in sozialen Beziehungen, transformativer Gerechtigkeit und den wissenschaftlichen Methoden.

Pragmatistische Wahrheit und induktives Schliessen

Ein Versuch der Rechtfertigung induktiver Schlüsse ausgehend vom Wahrheitsverständnis von William James

1. Einführung und Überblick

Der schwarze Schwan hat selbst nichts dazu beigetragen, zur Redewendung und zum Symbol der Wissenschaftstheorie geworden zu sein. Dass er sich im erstgenannten Sinn als „unvorhersehbares Ereignis" (Scheller 2020) hervorgetan hat, ist allein dem Umstand geschuldet, dass er zur falschen Zeit am falschen Ort war – fern von den Naturforschenden, die eines Tages zur Überzeugung gelangten, Schwäne seien ausnahmslos weiss. War dieser Schluss zu voreilig? Waren noch nicht genügend Beobachtungen angestellt worden? Oder war das Vorgehen der Forschenden gar unwissenschaftlich?

Das Induktionsproblem, das anhand dieses historischen Beispiels illustriert wird, wurde bereits vor fast 300 Jahren von David Hume thematisiert. Auf seine Infragestellung der Legitimität induktiver Schlüsse wird nach wie vor verwiesen, wenn sich die philosophische Debatte um ihren Stellenwert in der Wissenschaft von neuem entfacht. Hume erscheint beispielsweise die „Folgerung […] keineswegs notwendig", dass Brot uns zukünftig ernähren wird, nur weil es das in der Vergangenheit getan hat ([1748] 1964, 44). Vielmehr ist ihm zufolge das „Gegenteil jeder Tatsache […] immer möglich" und die Aussage, dass die Sonne morgen nicht aufgehen werde, ist „ein nicht minder verständlicher Satz und nicht widerspruchsvoller als die Behauptung, daß sie aufgehen wird" ([1748] 1964, 35–36). Wenn Ulrich Will aus diesem Anlass schreibt, dass es *a priori* nicht eher zu erwarten ist, „[d]aß sich die Billardkugel nach dem Stoß durch eine andere fortbewegt […], als daß sie in Ruhe verharrt oder sich in einen Elefanten verwandelt", (1981, 9) erscheint uns das zwar intuitiv absurd, aber gibt uns gleichzeitig zu verstehen, was mit dem Induktionsproblem auf dem Spiel stehen dürfte: Die alltägliche Anwendung von induktiven Schlüssen – sei es in der Form von Gewohnheiten, als bewusste Wahl von in der Vergangenheit für zielführend befundenen Verhaltensweisen oder als „Wissen über die Welt", das wir uns vielleicht zuschreiben würden – steht in drastischem Widerspruch zu der Schwierigkeit ihrer Rechtfertigung. Selbst wenn wir unsere Erwartungen an die Zukunft legitimieren möchten, indem wir argumentieren, dass uns frühere Erfahrungen schon oft erfolgreich die Zukunft gedeutet haben, leisten wir keine Rechtfertigung für diese Denkart, sondern bilden nur ein weiteres induktives Argument.

Es wäre dabei mehr als zu viel gesagt, dass sich in diesem Gebiet der Philosophie über die Jahre eine Rechtfertigung durchgesetzt habe, die erklärt, warum wir beispielsweise von einem noch nicht ausgegrabenen Smaragd erwarten dürften, dass er ebenso grün ist wie die bisher gefundenen Edelsteine dieser Art. Ganz im Gegenteil zeigt Goodman, dass selbst die vorsichtige Formulierung, wonach sämtliche bisher gefundenen Smaragde durch ihre Gemeinsamkeit, grün zu sein, die Aussage „alle Smaragde sind grün" immerhin *stütze*, irreführend ist: Es ist genauso gut denkbar, dass die grüne Farbe der gefundenen Edelsteine Zeugnis ihrer Eigenart ist, nur bis zu einem gewissen Zeitpunkt grün und von da an rot zu sein – eine durchaus denkbare Eigenschaft, die Goodman in seinem Gedankenexperiment „grot" nennt (1988, 98–99). Ob also alle Smaragde grün oder grot seien, ist auf Basis der vorliegenden Beobachtungen scheinbar nicht zu beurteilen. Keine der beiden Möglichkeiten ist der anderen vorzuziehen, während sich jeweils ihre Vorhersagen widersprechen. Selbst wenn also bisherige Beobachtungen Aussagen über die Zukunft legitimieren *könnten*, müssten wir uns dennoch zugestehen, dass wir nicht wissen können, von welcher Eigenschaft eine festgestellte Gleichartigkeit bisheriger Beobachtungen tatsächlich Zeugnis ist. Vor dieser Ausganslage erstaunt es

schliesslich wenig, dass Karl Popper der verheissungsvollen Verkündigung, er habe möglicherweise eine Lösung für das Induktionsproblem gefunden (1973, 13), die Ausformulierung einer Wissenschaftstheorie folgt, die gänzlich ohne induktive Schlüsse auskommen soll. Den Preis für seine Theorie – der Verzicht auf die Begriffe „Wissen", „Wahrheit" und „Wahrscheinlichkeit" (vgl. Popper [1935] 1994, 223) – sind nicht alle bereit zu bezahlen. Dazu kommt, dass es scheinbar viele nicht als Lösung des Induktionsproblems gelten lassen, wenn keine Rechtfertigung für induktive Schlüsse geliefert wird (vgl. Popper 1973, 40), vermutlich auch aufgrund der Auffassung, dass induktives Schliessen letztlich doch Bestandteil wissenschaftlicher Praxis ist. Sollte dies so sein, wird angesichts des wissenschaftlichen Fortschritts und der parallel dazu geführten philosophischen Streitigkeiten die Frage aufgeworfen, ob die Thematik der Induktion – „the glory of science and the scandal of philosophy" (Broad 1926, 67) – nicht doch eher etwas über die Eigenarten der Philosophie zu verstehen gibt und weniger über den Sinn wissenschaftlicher Praxis.

Popper, dessen Wissenschaftstheorie in der folgenden Untersuchung als Ausgangspunkt dient, hat mit seiner Arbeit einen nicht zu übersehenden Beitrag für diese philosophische Disziplin geleistet. Seine betonte Vorsicht, induktive Schlüsse als Methode aussenvor zu lassen und der Anspruch, möglichst unfehlbare und rein rationale Vorgehensweisen allein als wissenschaftlich zu bezeichnen, bringt allerdings einige Probleme mit sich. Die Hauptfunktion der Wissenschaft besteht demnach nur darin, die Falschheit von Theorien methodisch zweifelhaften Ursprungs[1] empirisch nachzuweisen. Nicht falsifizierte Theorien erlangen dabei zwar „Bewährtheit", können aber niemals für wahr befunden werden, da eine spätere Falsifikation zu keinem Zeitpunkt auszuschliessen ist. Es ist diese Konsequenz, sich von den Begriffen der Wahrheit und des Wissens verabschieden zu müssen, die hier zum Rückgriff auf einen alternativen Wahrheitsbegriff – demjenigen des Pragmatismus – motiviert. Daran anknüpfend möchte ich untersuchen, ob die Denkart, wie sie zum pragmatistischen Wahrheitsverständnis geführt hat, für die Debatte um das Induktionsproblem fruchtbar gemacht werden kann. Hierzu sind in einem nächsten Teil vorerst die Diskussion des Induktionsbegriffs und des Wahrheitsverständnisses Poppers erforderlich.

1 Der Prozess der Theoriebildung ist in Poppers Falsifikationismus fragwürdig, da er entweder beliebig oder doch das Resultat eines Induktionsschlusses ist. Dieser Umstand wird im späteren Verlauf der Untersuchung ausgeführt.

2. Unterschiedliche Induktionsbegriffe

Der Begriff „induktiver Schluss" wird nicht immer einheitlich verwendet. Manchmal wird damit ausschliesslich auf die Form eines enumerativen Induktionsschlusses verwiesen. Hume, der oft als Ausgangspunkt der Debatte betrachtet wird, bringt Beispiele dieser Art, wenn er etwa über die künftige Wirkung von Brot auf den menschlichen Körper oder den morgigen Sonnenaufgang nachdenkt. Bei der enumerativen Induktion handelt es sich um eine Schlussart, bei der wiederholte Einzelbeobachtungen zum Anlass genommen werden, Aussagen über Unbeobachtetes zu tätigen, insofern angenommen wird, dass das noch Unbeobachtete den bisherigen Beobachtungen in gewisser Weise entsprechen wird. Dabei spielt es genaugenommen keine Rolle, ob die durch den Induktionsschluss angenommenen Phänomene zum Zeitpunkt des Schlusses bereits existieren – wie etwa die Farbe von noch nicht ausgegrabenen Edelsteinen – oder erst eintreten werden – wie die Energie, die mir das Frühstücksbrot von morgen womöglich liefern wird. Gemeinsam ist diesen Möglichkeiten, dass die Folgen, die der Induktionsschluss nahelegt, ausserhalb der Erfahrung der Schliessenden liegen.

Wie weit der Begriff „Induktion" allerdings auch aufgefasst werden kann, zeigt etwa die Liste mit vierzehn verschiedenen Formen induktiver Schlüsse von Da Costa und French, die selbst damit nicht etwa Vollständigkeit beanspruchen. Hier finden sich neben der enumerativen Induktion auch statistische Schlussarten wie beispielsweise die von einer Stichprobe auf die Grundgesamtheit oder auf einen weiteren Einzelfall. Zudem fallen für diese Autoren auch Analogieschlüsse und Autoritätsargumente unter diesen Begriff (vgl. Da Costa und French 1989, 339–340). Einen solchen „abstrakteren Induktionsbegriff" umreisst Will in Abgrenzung zum Begriff der Deduktion folgendermassen:

> „Deduktive Argumente sind erstens durch eine logische Folgebeziehung gekennzeichnet, kraft deren sich die Wahrheit von den Prämissen auf die Konklusion überträgt, wenn die ersteren wahr sind. Sie sind notwendig wahrheitskonservierend. Zweitens führt bei ihnen der Gehalt der Konklusion nicht über den der Prämissen hinaus, sie sind mithin nicht gehaltserweiternd. Induktive Argumente kennzeichnet es demgegenüber, daß sie nicht notwendig wahrheitskonservierend, andererseits aber gehaltserweiternd sind.
>
> Drittens charakterisiert sie die ihnen eigentümliche logische Beziehung zwischen Prämis-

> sen und Konklusion, die je nach der zugrundeliegenden Induktionstheorie durch eine andere Formulierung angemessen zum Ausdruck gebracht erscheint. Die Prämissen bestätigen die Konklusion, machen sie wahrscheinlich, stützen sie, verleihen ihr Glaubwürdigkeit oder rechtfertigen einen rationalen Glaubensgrad in bezug auf sie." (Will 1981, 3)

Es sind also drei grundlegende Charakteristiken, die ein induktives Argument von einem deduktiven unterscheiden: Induktionsschlüsse sind gehaltserweiternd, nicht wahrheitskonservierend[2] und weisen keine logische Folgebeziehung zwischen Prämissen und Konklusion auf. Aus dieser Perspektive lässt sich das Grundproblem der Debatte um die Induktion präzise ausdrücken: Durch die Induktion lassen sich keine wahrheitskonservierenden Argumente bilden, die über getätigte Beobachtungen hinausgehen und in diesem Sinne gehaltserweiternd sind. Gehaltserweiternde Argumente können nicht in gleicher Weise gerechtfertigt werden, wie wahrheitskonservierende Argumente. Wenn die Wissenschaft gehaltserweiternd argumentieren will oder soll, kann sie dies nicht mit der gleichen Sicherheit tun, mit der wahrheitskonservierende Schlüsse gezogen werden. Es erscheint in gewissem Sinne trivial: Über Unbeobachtetes kann unmöglich mit dem gleichen Wahrheitsanspruch ausgesagt werden wie über Beobachtetes. Wer demnach induktive Schlüsse an der Sicherheit misst, mit der deduktiv geschlossen werden kann, stellt offensichtlich zu hohe Anforderungen.

3. Poppers Wahrheitsverständnis und dessen Konsequenzen für die Induktion

Mir scheint, dass Popper ein in der Wissenschaft verbreitetes Weltbild vertritt, demzufolge es eine objektive, von uns unabhängige Realität gibt. Die Wissenschaft, für die nach Popper „der Wahrheitstrieb […] wohl stärkste[r] Antrieb" ist, scheint sich Zugang zu dieser Realität verschaffen zu wollen, „[o]bwohl Wahrheit und Wahrscheinlichkeit für sie unerreichbar ist" ([1935] 1994, 223). Eine Theorie hat ihm zufolge Wahrheitsanspruch, solange sie nicht falsifiziert ist – wohl aber überprüft werden kann (vgl. Popper [1935] 1994, 14) – und gegebenenfalls, dass sie bereits mehreren Überprüfungen standgehalten hat, einen höheren Bewährungsgrad (vgl. Popper 1973, 32). Es soll hier nicht diskutiert werden, ob Popper mit seinem Konzept der „Bewährtheit" von Theorien letztlich doch gewisse induktive Schlüsse zulässt, sondern erschlossen werden, wie Popper über die Welt, Erkenntnismöglichkeiten und Wahrheit denkt. Wenn ich ihn richtig verstehe, ist eine Theorie dann wahr, wenn sie mit der objektiven Realität übereinstimmt, was sich gegebenenfalls so ausdrückt, dass sie sämtlichen Überprüfungen standhalten und nie falsifiziert werden würde. Da die Falsifikationsversuche jedoch unerschöpflich sein dürften, ist der Punkt, an dem gesagt werden kann, dass eine Theorie „wahr" sei, unerreichbar. Zudem darf wohl nicht davon ausgegangen werden, dass eine Überprüfung, die nicht zur Falsifikation geführt hat, dies in Zukunft nicht doch bewirken könnte. Die Wissenschaft versucht sich also die objektive Realität verständlich zu machen, sich einen Zugang zu ihr zu verschaffen, ohne es mit Gewissheit erkennen zu können, falls sie ihr Ziel tatsächlich erreicht haben sollte. Man könnte in anderen Worten sagen, dass Nicht-Falsifikation bestenfalls daher rühren könnte, dass eine Theorie in gewissen Aspekten mit der metaphysischen Beschaffenheit der Welt ausreichend übereinstimmt, wobei epistemische Einschränkungen eine Feststellung dieses Umstands verhindern.

Wenn wir dieses Verständnis von Wahrheit und den erarbeiteten weiten Induktionsbegriff mit einem Beispiel Musgraves (vgl. 2004, 23) auf den Bereich des Alltäglichen anwenden, stellen wir fest, dass Schwierigkeiten allgegenwärtig sind. Habe ich zum Beispiel die Erfahrung, dass vor mir ein Tisch sei, wäre es bereits ein gehaltserweiternder Schluss, daraus abzuleiten, dass vor mir tatsächlich ein Tisch ist. Eine Erfahrungsaussage hat keineswegs eine Aussage über die objektive Realität zur logischen Folge und dementsprechend würde es sich hier um einen Induktionsschluss nach dem weiten Begriffsverständnis handeln, denn es wird aufgrund einer Erfahrung auf etwas Unbeobachtetes – die objektive Realität des Tisches – geschlossen. Wenn wir diesen Schluss für unzulässig halten und stattdessen mit Popper die Existenz des Tisches zur Hypothese und sinnliche Wahrnehmungen des Tisches zu prüfbaren Folgen erklären, so würde der Wahrheitsbegriff auch aus dem Alltag verschwinden müssen.

Man kann Popper wohl nicht unterstellen, dass er so über alltägliche Erkenntnis denkt, obgleich sich die Frage, ob und wie er eine Abgrenzung zwischen alltäglichen und wissenschaftlichen Gegenständen vornehmen würde, trotzdem stellt. Mir scheint jedenfalls,

2 Es wird hier angenommen, dass sich Wills Wendung „nicht notwendig wahrheitskonservierend" darauf bezieht, dass man für eine wahre Konklusion eines induktiven Schlusses argumentieren *könnte*, aber eine Wahrheitskonservierung rein formal nicht vorliegt.

dass wir bereits bei solchen basalen Beispielen viel eher sagen würden, dass die Erfahrung den Schluss auf die Realität gestattet, als dass sie die Überprüfung einer Hypothese wäre – zumal im Alltag die Überprüfung zeitlich der Formulierung einer Hypothese vorausgehen müsste. Ungeachtet des Geltungsanspruchs von Poppers Theorie für Alltagserfahrungen scheinen wir demnach bereits in diesem Bereich Schlüsse zu ziehen, die man strenggenommen für induktiv halten dürfte – was natürlich diese Vorgehensweise nicht etwa legitimiert. Wer allerdings den Wahrheitsbegriff so verwenden will, dass man ihn an den Zustand vollständiger Übereinstimmung einer Vorstellung oder Theorie mit der objektiven Realität koppelt – wie er gegebenenfalls noch nicht mal festgestellt werden könnte –, stellt meines Erachtens zu hohe Anforderungen an die Wahrheit.

Wie im Zuge der Diskussion um den Induktionsbegriff erarbeitet wurde, scheinen sich die beiden Eigenschaften von Argumenten, wahrheitskonservierend beziehungsweise gehaltserweiternd zu sein, gegenüberzustehen. Die Lösung des Dilemmas wäre eine Schlussform, die beide Kriterien erfüllt und insofern die Wahrheit der Prämissen auf eine gehaltserweiternde Konklusion beziehungsweise die Wahrheit bisheriger Erfahrungen auf Unerfahrenes übertragen könnte. In Berücksichtigung des Wahrheitsbegriffs, wie er oben diskutiert wurde, kann es bereits bei der Frage nach der Wahrheit von Prämissen zu Schwierigkeiten kommen. Popper umgeht diese, indem er lediglich aus Theorien abgeleitete Konsequenzen überprüft, die als deduktive Ableitungen genau so wahr sein müssen wie die Theorie, aus der sie abgeleitet werden, wobei die Prüfungen höchstens zeigen, in welcher Weise die Theorie nicht für falsch erklärt werden kann. Somit wird nichts positiv zur Wahrheit einer Theorie beigetragen. Wollte man aber für die Berechtigung induktiver Schlüsse argumentieren, so scheint es, kann man sich nicht auf eine logisch notwendige Wahrheitskonservierung verlassen und man müsste der Frage nach Wahrheit ausweichen. Falls ein induktives Argument zu einer wahren Konklusion führen würde, so wäre der Grund dafür nicht primär in ihrer logischen Beziehung zu den Prämissen zu suchen, weil die Gehaltserweiterung und somit die Konklusion unterschiedlich ausfallen kann, ohne die logische Form des Arguments zu verändern – oder wie Goodman schreibt: „Die Bestätigung einer Hypothese durch einen Einzelfall hängt stark von Eigenschaften der Hypothese ab, die nichts mit ihrer syntaktischen Beschaffenheit zu tun haben“ (1988, 97).

4. Der pragmatistische Wahrheitsbegriff

Da es also fraglich ist, ob Poppers Wahrheitsverständnis überhaupt sinnvoll auf mögliche Rechtfertigungen von Induktionsschlüssen angewendet werden kann, und das Verschwinden der Wahrheit aus der Wissenschaft ohnehin zur Suche nach alternativem Wahrheitsverständnis motiviert, soll nun auf eine Strömung Bezug genommen werden, die es zu ihrem Ziel erklärt hat, „philosophische Streitigkeiten zu schlichten, die sonst endlos wären“ (James 1908, 27). Dieser Anspruch, den Versuch zu wagen, sich einen neuartigen Zugang zur Wahrheit zu verschaffen, hat schliesslich zu starker Kritik geführt, was Andreas Graeser in der mangelnden Rücksicht auf besagten Anspruch begründet sieht. William James, der Urheber des pragmatistischen Wahrheitsverständnisses, wollte gemäss Graeser primär „jene Motive […] eruieren, die uns normalerweise dazu bestimmen, Urteile, Vorstellungen usw. als wahr bzw. als falsch anzusehen“ (Graeser 2000, 278). James argumentiert im Zuge der Entwicklung seiner Theorie gegen die Position der „Rationalisten“ beziehungsweise „Intellektualisten“. Auch Popper könnte zumindest teilweise der von James kritisierten Position zugerechnet werden, insbesondere weil er die Wahrheit einer Theorie zwar als Möglichkeit akzeptiert, diese aber – in James‘ Worten zum rationalistischen Verständnis „über den Kopf der Erfahrung hinweg“ (1908, 138) – an gänzlich objektiven Kriterien festmacht und ihre Erkenntnis und folglich auch Wissen ausschliesst. Ich möchte versuchen, das pragmatistische Wahrheitsverständnis einigermassen kompakt wiederzugeben, was durch seine eigentümliche Beschaffenheit nicht allzu leichtfallen dürfte und zumindest einer anschliessenden Interpretation bedürfen wird, die es für das Nachfolgende anwendbar machen soll.

James stimmt mit seiner Gegenposition darin überein, dass sich wahre Vorstellungen dadurch auszeichnen, dass sie mit der Wirklichkeit übereinstimmen, wobei er jedoch unter „Übereinstimmung“ und „Wirklichkeit“ etwas anderes versteht als die Rationalisten (vgl. 1908, 124). Statt des „absolut leere[n], rein statische[n]“ (James 1908, 44) Verständnisses von Übereinstimmung, das nach Popper, wenn überhaupt, eher zufällig entstehen würde, plädiert er für eine dynamische Auffassung. Für James ist Übereinstimmung mit der Wirklichkeit ein „Vorgang des Geführt-Werdens“ und könne „im weitesten Sinne nichts anderes heißen, als zu dieser Wirklichkeit oder in ihre Umgebung geradeaus hingeführt werden oder mit derselben in eine derartige wirksame Berührung gebracht werden“ (James 1908, 134). Diese zugegebenermassen diffuse Beschreibung lässt zumindest vermuten, dass James Übereinstimmung mit der Wirklichkeit gewissermassen als Interaktion mit ihr versteht. Es fällt zudem auf, dass

James nie von *der*, sondern stets von *einer* Wirklichkeit schreibt, beziehungsweise in konkreten Fällen von einer damit übereinstimmenden Vorstellung von *dieser* Wirklichkeit (1908, 134). Ganz anders als beim Rationalismus ist hier nicht von einer objektiven Realität die Rede, die epistemisch mehr oder weniger zugänglich ist. Im pragmatistischen Verständnis ist „Wirklichkeit so viel […] als erfahrbare Wirklichkeit" und „sie selbst und die wahren Erkenntnisse, die die Menschen von ihr gewinnen, [sind] in einem fortwährenden Veränderungsprozeß begriffen" (James 19808, 142). Eine konkrete Wahrheit scheint sich also aus der Interaktion mit einer konkreten Wirklichkeit, einer Erfahrung, zu ergeben, wobei sich Wahrheit und Wirklichkeit in wechselseitiger Beziehung verändern können:

> „Aus Tatsachen ergeben sich Wahrheiten, diese aber dringen wieder weiter in die Tatsachen ein und fügen neue hinzu. Diese neuen Tatsachen schaffen oder offenbaren neue Wahrheiten und so geht es immer weiter bis ins Unendliche. Die Tatsachen selbst sind dabei nicht wahr, sie sind einfach." (James 1908, 142–143)

Dieses „Hinzufügen neuer Tatsachen" klingt zunächst ziemlich mysteriös. Jedoch scheint dies James' Idee von veränderbarer Wirklichkeit zu entsprechen, insofern durch das Geführt-Werden neue Erfahrungen gemacht werden können, was mit einer verbreiteten Auffassung von wissenschaftlichem Fortschritt gewiss einhergeht. Die bisher angesprochenen Begriffe „Wahrheit", „Wirklichkeit" und „Übereinstimmung" scheinen demnach eng verflochten zu sein. Zu diesem wechselseitigen Verhältnis kommen allerdings noch zwei weitere Begriffe hinzu, die James im Zuge der Erläuterung mehrfach verwendet und von denen besonders letzterer für grosses Unverständnis gesorgt hat: Verifikation und Nützlichkeit. Das interaktionistische Verständnis von Übereinstimmung mit der Wirklichkeit scheint als Prozess noch spezifischer mit „Verifikation" beschrieben zu werden: Dass Wahrheit für Vorstellungen nämlich ein „Vorkommnis" sei und dass sie „durch Ereignisse wahr gemacht" werde – ein Geschehen, das als „Vorgang [der] Selbstbewahrheitung" (James 1908, 126) zu verstehen sei – ist James zufolge der Prozess, den man Verifikation nennt. Meiner Einschätzung nach handelt es sich bei einem solchen Vorgang um das Schaffen von Übereinstimmung zwischen Vorstellung und Wirklichkeit. Demgegenüber haben sich angesichts des Begriffs der Nützlichkeit, der nach der pragmatistischen Auffassung wie die anderen Begriffe stets mit der Wahrheit einer Vorstellung einhergehe, verständlicherweise viele kritische Stimmen erhoben. Unnütze Wahrheiten sind zweifellos eine Alltagserfahrung und wenn ich beispielsweise zu spät einsehe, dass ich mich zur Erreichung eines Ziels anders hätte verhalten sollen, so scheine ich somit eine überaus unnützliche Wahrheit vor mir zu haben. Es klingt als dürften wir nach pragmatistischer Auffassung jeweils diejenige Vorstellung für wahr halten, die uns gerade nützlich erscheint. Ich plädiere jedoch für ein Begriffsverständnis, über das die Bezeichnungen „Brauchbarkeit" oder gar „Anwendbarkeit" besser Aufschluss geben. Der oben zitierte Passus, der den Vorgang des Erkenntnisfortschritts formuliert, scheint eine mögliche Folge einer in diesem Sinne brauchbaren Vorstellung aufzuzeigen. Eine Vorstellung ist brauchbar hinsichtlich ihres Interaktionspotenzials mit einer Wirklichkeit, geht aus Erfahrung dieser Wirklichkeit hervor und bewirkt weitere Erfahrung. Der Zusammenhang der Wahrheit und der Brauchbarkeit einer Vorstellung ist James zufolge bikonditional und bedeute „daß hier ein Gedanke da ist, der verwirklicht und verifiziert werden kann" (1908, 128). In diesem engen Zusammenhang muss man sich diese Begriffe dem Pragmatismus zufolge denken.

> „Die Erfahrung ist in fortwährender Veränderung, und unsere Kenntnisnahmen der Wahrheit sind als psychische Prozesse ebenfalls in steter Veränderung begriffen. So viel wird der Rationalismus zugeben; aber niemals wird er zugeben, daß die Wirklichkeit selbst und daß die Wahrheit selbst veränderlich ist." (James 1908, 143)

Ich bin offenbar rationalistisch genug, um diese Aussagen für die beträchtlichsten des pragmatistischen Wahrheitsverständnisses zu halten. Es sind die zentralen Punkte, die James von Auffassungen wie derjenigen Poppers abgrenzen. Damit diese Eigenart angemessen verstanden wird, ist es unerlässlich, zu berücksichtigen, dass James nicht ausschliesst, dass es „[e]in absolut Wahres in dem Sinne, daß keine künftige Erfahrung es ändern kann" (1908, 141) geben könnte. Es ist somit nicht gesagt, dass es keine objektive Realität als Gegenstand der metaphysischen Dimension gibt, sondern dass diese für unsere Erfahrung und damit auch für unsere Wahrheit – als Gegenstände der epistemischen Dimension – unerheblich ist. Die Möglichkeit der Wahrheit als eine Übereinstimmung mit der Wirklichkeit basiert demnach auf der Unterscheidung von epistemischer und metaphysischer Dimension und in diesem Sinne sollen hier die zur Unterscheidung angewendeten Begriffe „erfahrbare Wirklichkeit" und „objektive Realität" eingeführt

werden. Ich möchte dafür argumentieren, dass James gute Gründe für seine Kritik an gewissen Vorstellungen von Wahrheit hat. Der Begriff der Übereinstimmung ist im sogenannten rationalistischen Verständnis tatsächlich leer und unter Umständen sogar eher zufällig. Es stellt sich die Frage, wie ein Gedanke, eine Vorstellung oder eine Theorie überhaupt zu ihrem Inhalt kommen soll, wenn sie zwar mit der objektiven Realität übereinstimmen können soll, aber nicht aus einer Interaktion mit der erfahrbaren Wirklichkeit hervorgeht. Ich würde nicht behaupten, dass sich James sträuben würde, eine Vorstellung, wenn sie tatsächlich mit der objektiven Realität übereinstimmen sollte, als wahr anzuerkennen. Nur gibt es zum Zeitpunkt zu dem derartiges von einer Vorstellung behauptet wird für gewöhnlich nichts, was diese Übereinstimmung gewährleistet; keine Interaktion zwischen einer derartigen Realität und einer Vorstellung, keine Verifikation, die eine Vorstellung bewahrheiten und ein angemessenes Verhältnis zwischen Vorstellung und Realität bewirken würde. Der Ansatz wäre nach James deshalb unzulässig, weil ihm zufolge Wahrheit nicht „vor den Tatsachen“ (1908, 140) existiert und erst durch Erfahrung möglich ist, weswegen das Ideal des „absolut Wahren“ zwangsläufig „gleichen Schritt mit dem vollkommen weisen Mann“ (1908, 141) halten muss. Da jedoch auch ein solches Ideal nur für eine erfahrbare Wirklichkeit gehalten werden würde, von der unmöglich gesagt werden kann, dass zukünftige Erfahrung sie nicht mehr verändere, kann eine solche Wahrheit nicht als die endgültige erkannt werden. Ich sehe nicht, was der Pragmatismus dieser Ansicht Poppers entgegenzusetzen hätte. Eine angenehme Folge der pragmatistischen Auffassung ist nun – denken wir an den vollkommen weisen Mann –, dass nicht wie bei Popper, der Wissen völlig ausschliesst, Wissen sozusagen die Konsequenz von Wahrheit ist. Durch die Vorgänge, die zur Erzeugung von Wahrheit erforderlich sind, wird zwangsläufig auch eine gerechtfertigte Überzeugung bewirkt. So verheissungsvoll diese ganzen Entwicklungen auch sind, müssen sie natürlich auch wieder relativiert werden: Wahrheit und Wissen ist immer nur innerhalb gewisser Erfahrungsgrenzen möglich. Somit ist gegenüber der „absoluten“ strenggenommen stets von einer „relativen Wahrheit“ zu sprechen, die bereits morgen als falsch bezeichnet werden könnte, wenn man sie „absolut betrachtet“ (James 1908, 141). Wir können folglich heute etwas wissen, das wir morgen nicht mehr wissen können, ohne den Gehalt dieses ehemaligen Wissens zu verändern – dies ist der Fall, wenn sich die Erfahrungsgrenzen verschoben haben und sich somit die erfahrbare Wirklichkeit verändert hat.

5. Pragmatismus und Wissenschaftstheorie

Was sogleich auffällt ist, dass James mit dem wohl nicht allseits beliebten Charakteristikum seines Wahrheitsbegriffes eine Möglichkeit schafft, mit der Konklusion der pessimistischen Metainduktion umzugehen, wie sie beispielsweise von Fahrbach diskutiert wird (vgl. 2017, 5042). Dass auch heutige etablierte Theorien dereinst überholt sein werden, wie es in der Geschichte schon oftmals passiert ist, wäre lediglich der Veränderung der für die Theorie relevanten Wirklichkeit geschuldet. Eine andere Frage ist die nach der Methode, mit der eine Theorie mit ihrer Wirklichkeit interagiert, beziehungsweise was aus der erfahrbaren Wirklichkeit gemacht wird. Es ist nicht gesagt, dass sich Theorien nicht qualitativ unterscheiden können, in der Weise, wie sie die gegebenen Umstände nutzen. Zweifellos ist die während der Theoriebildung zugängliche Wirklichkeit auch von der Qualität des wissenschaftlichen Diskurses abhängig, insofern sämtliche geltende Wahrheiten „aufgespeichert und für jedermann verwendbar gemacht“ (James 1908, 134) werden, sodass eine kohärente Menge von „Sinneserfahrung[en], die irgendjemand in seiner Vorstellung abgebildet hat“ (1908, 136), zustande kommt. James denkt im Zuge der Entwicklung seines Wahrheitsbegriffs über Anforderungen an eine solche Theorie nach:

> „Wir müssen eine Theorie finden, mit der wir arbeiten können, und das ist etwas ungemein Schwieriges, denn unsere Theorie muß zwischen allen früheren Wahrheiten und einigen neuen Erfahrungen vermitteln. Sie darf dem gesunden Menschenverstand und den früheren Überzeugungen so wenig als möglich zuwider laufen, und sie muß dabei zu etwas Anschaulichem hinführen, zu etwas, das in exakter Weise verifiziert werden kann. Beides ist nötig, damit eine Theorie ‚arbeite‘, und so sind wir dabei so in die Enge getrieben, daß für jede Hypothese nur sehr wenig Spielraum bleibt.“ (James 1908, 136)

Es sind also vorwiegend zwei Aspekte, die James bei der Theoriefindung ins Zentrum rückt. Einerseits sollen die neuen Erfahrungen kohärent mit bestehenden Wahrheiten abgestimmt werden. Hier kann es gegebenenfalls um die Frage gehen, welche etablierten Überzeugungen verworfen werden müssen oder sollen, beziehungsweise wie weit eine Erklärung für neue Erfahrungen bestehende Wahrheiten aufrechterhalten soll. Dass die Theorie andererseits zu etwas „Anschaulichem hinführen“ müsse, ist innerhalb von James‘ Wahrheitsver-

ständnis eine absehbare Forderung. Dabei teilt er diese Bedingung für die Wahrheitsfähigkeit einer Theorie mit Popper, der jedoch offensichtlich nicht von Verifikation sprechen würde. Die von James angedachten Verifikationen erinnern hingegen ein wenig an Poppers Konzept von Bewährtheit, das sich ohnehin in pragmatistisches Vokabular übersetzen liesse. Wenn aus Theorien abgeleitete Konsequenzen im Zuge ihrer Prüfung nicht falsifiziert werden, kann das als Übereinstimmung verstanden werden, insofern es sich um eine wirksame Berührung mit der Wirklichkeit handelt – oder als Brauchbarkeit der Theorie, die durch die wahrmachenden Ereignisse zu neuen Erfahrungen führt. Mit James' Auffassung lässt sich meines Erachtens gar besser nachvollziehen, wie Popper dem Vorwurf, durch seinen Bewährtheitsbegriff gewisse induktive Schlüsse zuzulassen, entgegentreten möchte: Wenn Bewährtheit, wie Popper schreibt, „ein Bericht über bisherige Leistungen" einer Theorie sein soll, die aber „nicht das geringste über [ihre] zukünftigen Leistungen oder […] ‚Verläßlichkeit'" (1973, 30) aussagen könne, kann Bewährtheit als Qualitätsindikator der Übereinstimmung mit einer konkreten erfahrbaren Wirklichkeit verstanden werden, wobei eine der Theorie widersprechende Veränderung der Wirklichkeit stets möglich bleibt.

6. Verifikation und Falsifikation

Wenn wir die Auffassungen von James und Popper einander gegenüberstellen, ist schnell ersichtlich, dass jeweils ein zentraler Begriff für den Umgang mit Vorstellungen und Theorien beim jeweils anderen Denker kaum auftritt: So sieht Popper die Falsifikation als alleiniges echtes Ereignis, das im Zusammenhang mit Theorien auftreten kann, da „Verifikation" als Synonym für „Bewahrheitung" aus erläuterten Gründen nicht in Frage kommt. James andererseits erwähnt zwar „falsche Vorstellungen" als solche, die nicht mit einer erfahrbaren Wirklichkeit in Interaktion treten können, interessiert sich jedoch viel eher für die „Verifikation" wahrer Vorstellungen. Falsifikation ist für James ein Vorgang von anderer Qualität: Dadurch, dass neue Erfahrungen, die einer Vorstellung widersprechen, hinzukommen, stimmt diese nicht mehr mit der erfahrbaren Wirklichkeit überein. Ich halte das für etwas anderes als die Einsicht, mit einer Theorie auf dem falschen Weg zu sein.[3] Insofern das Hinzukommen neuer Erfahrungen viel eher eine Erweiterung als eine blosse Veränderung der Wirklichkeit ist, tritt mit Falsifikation stets auch Fortschritt auf. Neue Erfahrungen ersetzen nicht einfach alte und verschieben die Wirklichkeit nicht einfach mit gleichbleibendem Spektrum in eine gewisse Richtung: Selbst wenn Wahrheiten verworfen werden müssen, ist die Frage danach, was zu ihrer Entstehung geführt hat und inwiefern sie gemessen an der neuen Wirklichkeit fehlerhaft sind, sinnvoll und wichtig, da ihre Beantwortung eine Erkenntniserweiterung bedeutet. Poppers Anspruch, jede noch so bewährte Theorie ohne Weiteres fallen zu lassen, sollte eine weitere Überprüfung zur Falsifikation führen, halte ich dahingehend für fragwürdig. Stellt sich nicht gerade da die Frage, wie es kommen konnte, dass eine „absolut betrachtet" falsche Theorie erfolgreiche Voraussagen getroffen hat? Mir scheint genau darin der wissenschaftliche Fortschritt zu liegen, dass Falsifikation oftmals eine Erweiterung unserer Kenntnisse bewirkt. Das von Popper angeführte Beispiel, wonach 1931, als Wasser, Sauerstoff und Wasserstoff die am besten bekannten Stoffe waren, Deuterium und schweres Wasser entdeckt wurden, sodass das bisherige Verständnis dieser Substanzen überholt war, (vgl. 1973, 22) macht diese Charakteristik meines Erachtens deutlich: Nur das bis 1931 erarbeitete Verständnis dieser Stoffe – wenn dieses nicht auch überhaupt erst die neuen Erfahrungen ermöglicht hat – konnte dazu führen, dass die neuen Erkenntnisse in ein neues und besseres Verständnis eingebettet werden konnten. Die starke Überzeugung, mit der gemäss Popper das bis dahin vorhandene Verständnis vertreten wurde, könnte demnach als Zeichen der starken Übereinstimmung und Verifikation der alten Theorie mit der damals erfahrbaren Wirklichkeit im pragmatistischen Sinne verstanden werden. Es scheint mir unwahrscheinlich, dass solche starken Wahrheiten durch hinzukommende und dem bisherigen Verständnis widersprechende Erfahrungen für gänzlich nichtig erklärt werden müssten. Viel eher müssen die Gründe ihres Zustandekommens mit der Erneuerung in Kohärenz gebracht werden oder es muss ihnen zugestanden werden, dass sie erst zu den neuen Erfahrungen geführt haben. Es muss, wie James sagt, zwischen alten Wahrheiten und neuen Erfahrungen vermittelt werden, statt nach der Falsifikation einer Konsequenz einer Theorie mit leeren Händen und einem negierten Teilsatz von vorne zu beginnen. Denken wir uns ein Beispiel, das sich zwar wissenschaftshistorisch wohl nicht in dieser Weise zugetragen hat, aber dennoch diesen Punkt erläutert: Wenn in der Folge von Beobachtungen ein Gesetz angenommen wird, wonach die Dichte jeglicher Stoffe bei zunehmender Temperatur

3 An dieser Stelle sei noch angemerkt, dass ich die Begriffe „Vorstellung" und „Theorie", wie sie James beziehungsweise Popper vorzugsweise verwenden, nicht gesondert diskutiere, da ich ihre Unterscheidung in Bezug auf die diskutierten Punkte für unerheblich erachte. Ich glaube nicht, dass sich Poppers Aussagen fundamental von den hier zitierten unterscheiden würden, handelten sie von Vorstellungen statt von Theorien.

ab- und abnehmender Temperatur zunimmt und nach der Formulierung dieses Gesetzes (insofern wohl ein fiktives Beispiel) die Dichteanomalie von Wasser festgestellt wird, dann wäre es doch fahrlässig, das falsifizierte Gesetz vollständig fallenzulassen. Es erhebt sich dann die Frage, wie das Gesetz unter Berücksichtigung der neuen Erfahrung überhaupt zustande kommen konnte und warum es nun auf die einen Stoffe nicht und auf andere dennoch zutrifft. Nach der Erkenntnis der Dichteanomalie von Wasser würde – soweit ich das abschätzen kann – nicht nur das Wissen über Wasser erweitert, sondern zudem jenes über die Gültigkeit des Gesetzes, das auch danach noch auf einen Zusammenhang von Temperatur und Dichte verweist. Ferner dürfte durch die Fokusveränderung auf andere Stoffe bewirkt worden sein, dass diese nun in einer Weise betrachtet werden, die ohne die Falsifikation vielleicht nicht zustande gekommen wäre.

Es kann hierzu ein weiteres Argument angeführt werden, das verständnisübergreifend für eine Art höhere Qualität der Falsifikation gegenüber der Verifikation spricht. Es ist anzunehmen, dass Popper konkrete Erfahrungen auch im Zusammenhang mit einer jeweiligen erfahrbaren Wirklichkeit betrachten würde. Nur ist der Umstand, dass konkrete Wirklichkeiten jederzeit überholt werden können, für ihn Anlass, sie nicht als wahr zu bezeichnen. Dass stattdessen Falsifikation stets zulässig ist, impliziert, dass es nicht möglich ist, konkrete Konsequenzen einer Theorie in Bezug auf eine aktuelle Wirklichkeit zu falsifizieren, die in Bezug auf eine Erweiterung dieser Wirklichkeit nicht falsifiziert werden würden – sonst müsste Popper Falsifikation in gleicher Weise zurückweisen, wie Verifikation. Daraus würde folgen, dass Falsifikation über alle noch zu erreichenden wahrnehmbaren Wirklichkeiten hinweg etwas über die objektive Realität zu verstehen gibt. Da in diesem Sinne Falsifikation aussagen kann, was nie oder nie mehr für möglich erachtet werden wird, ist sie von höherer Qualität als die Verifikation, deren Gegenstand durch die Zeit nur relativen Geltungsanspruch erheben kann. Dennoch könnte man argumentieren, dass Popper einen Induktionsschluss zulässt, sobald er die einmalige Falsifikation eines Beobachtungssatzes durch eine konkrete empirische Prüfung für endgültig erachtet. Wenn hier nämlich nicht angenommen werden soll, dass derselbe Beobachtungssatz durch die gleiche empirische Prüfung auch in Zukunft falsifiziert werden würde, gäbe es keinen Grund dafür, diese Prüfung nicht zu wiederholen, denn in Zukunft könnte sie zu einem anderen Resultat führen. Es gäbe schliesslich genau so wenig Gründe dafür, eine Theorie zu verwerfen, weil sie durch eine Prüfung falsifiziert wurde, wie dafür, die Theorie für bestätigt zu halten, weil sich eine beobachtbare Konsequenz aus ihr ableiten lässt.

7. Die Notwendigkeit induktiver Schlüsse

Ich bin der Auffassung, dass Popper und James mit ihren jeweiligen Wahrheitsbegriffen als Pole eines grösseren Spektrums möglicher Positionen betrachtet werden können. Wie bei Popper epistemische Grenzen den Besitz von Wahrheit gänzlich verunmöglichen, sind es bei James genau diese epistemischen Grenzen innerhalb derer Ereignisse stattfinden, die als Wahrheiten zu verstehen sein sollen. Die wissenschaftstheoretischen Ansichten dieser zwei Denker scheinen mir dabei nicht grundverschieden zu sein, sondern in etwa auf folgende Gemeinsamkeit gebracht werden zu können: Wissenschaftliches Erkenntnisstreben ist der Versuch sich eine objektive Realität zugänglich zu machen, wobei jeweils kaum abzuschätzen ist, in welchem Verhältnis aktuelle Erkenntnis und objektive Realität stehen und ob wir Letzterer stetig oder überhaupt näherkommen. Worin sich die beiden Herangehensweisen grundlegend unterscheiden, ist der Bereich, den sie thematisieren. Während James, wie oben diskutiert, auch den Prozess der Theoriebildung miteinbezieht, belässt es Popper dabei, sich allein dem Umgang mit Theorien zu widmen. Warum dies so ist, deutet bereits sein Wahrheitsbegriff an: Wahrheit ist nicht zu erreichen, da es die absolute Gewissheit nicht geben kann. Mit diesem Anspruch wird folglich auch für den Falsifikationismus argumentiert, der seine Methode exakt begründen kann und sich jeglichen Vorwürfen, zu nicht ewig gültigen Überzeugungen führen, entziehen will. Dass demnach auch nichts Positives aus dem wissenschaftlichen Prozess gewonnen werden kann – von der umstrittenen Bewährtheit mal abgesehen – hat die Konsequenz, dass über die Theoriefindung geschwiegen wird. In keiner Weise liesse sich eine solche Methode zur Findung neuer Theorien mit denselben Ansprüchen rechtfertigen. Solange die Theorie prüfbare Konsequenzen ableiten lässt, scheint sie gut genug, der Popperschen Falsifikationsmethode unterzogen zu werden. Wie kommt allerdings eine solche Theorie zustande? Um eine Theorie mit erfahrbaren Konsequenzen zu bilden, braucht es einerseits Erfahrungen als Grundlage, da sie sonst höchstens als vollständiges Zufallsprodukt denkbar ist, und zudem dennoch von einem Subjekt hervorgebracht werden müsste, das zumindest irgendwelche Erfahrungen gemacht hat, die zumindest in geringer Weise ihren Einfluss haben dürften. Andererseits muss die Theorie über die ihr zugrundeliegende Erfahrung hinausgehen, sodass sie etwas Erfahrbares

impliziert, das so noch nicht erfahren wurde. In anderen Worten: Eine Theorie muss auf Basis von Beobachtetem Aussagen über Unbeobachtetes ermöglichen. Alles was nicht von dieser induktiven Art ist und dennoch Poppers Anforderungen genügt, wäre als Zufallsprodukt mindestens ebenso wenig zu rechtfertigen, wie die Induktion. Dies alles führt nun nicht auf eine Kritik an der Unvollständigkeit der Methode von Popper hinaus, sondern unterstreicht lediglich die Konsequenz aus seinem Anspruch an Rationalität: Popper thematisiert den Prozess der Theoriefindung nicht weiter, da kein derartiger, ausreichend rationaler Prozess denkbar ist, der seinen Ansprüchen genügen würde. Er will keine Induktion befürworten und muss es für seine Theorie auch nicht; aber seine Methode hat ohne Induktion nichts zu tun. Sobald wir uns zudem fragen, ob es nicht bessere und schlechtere Prozesse geben könnte, die zu einer falsifizierbaren Theorie führen – und mir scheint, dass die Antwort eine positive sein muss – haben wir hier eine Schwierigkeit. Dass sich Popper dieser Schwierigkeit aus guten Gründen entzieht, entschädigt jedoch nicht dafür, dass er dies – so hier mein Vorwurf – wider besseres Wissen tut.

8. Induktion bei James

James scheint keine grösseren Probleme damit zu haben, Anforderungen an einen Theoriefindungsprozess anzugeben, wie oben bereits diskutiert wurde. Ob dies auf andere Ansprüche oder den andersartigen Wahrheitsbegriff – der ebenso aus diesen hervorgehen dürfte – zurückzuführen ist, oder ob er das Induktionsproblem ohnehin nicht als solches auffasst, ist nicht sogleich zu beurteilen. Dass er zumindest einen eher leichtfertigen Umgang mit einigen induktiven Schlüssen zu haben scheint, gibt folgendes Zitat zu verstehen:

> „Ein weiterer wichtiger Grund dafür, daß wir im gewöhnlichen Leben auf vollständige Verifikation verzichten, ist, abgesehn von der Zeit-Ökonomie, auch der Umstand, daß die Dinge nicht in lauter Einzelexemplaren, sondern in Gattungen da sind.
>
> Unsere Welt hat ein für allemal diese Eigenschaft. Wenn wir also unsre Vorstellungen an einem bestimmten Exemplar einer Gattung unmittelbar verifiziert haben, so halten wir uns für berechtigt, sie ohne weitere Verifikation auf andere Exemplare anzuwenden. Ein Geist, der gewohnheitsmäßig die Gattung des Dinges, das er vor sich hat, erkennt und dann gemäß dem Gesetze der Gattung, ohne sich mit weiteren Verifikationen aufzuhalten, sofort zur Tat schreitet, wird in 99 von 100 Fällen ein ‚wahrer' Geist sein. Die Wahrheit seiner Urteile ist dadurch bewiesen, daß seine Handlungsweise die entsprechende ist und keine Widerlegung erfährt. (James 1908, 131)

Diese Passage wirft mehrere Fragen auf: Die Formulierung von James, wonach die Welt „ein für allemal" die Eigenschaft habe, aus Gattungen statt Einzelexemplaren zu bestehen, scheint seinem eigenen Wahrheitsbegriff zuwiderzulaufen, insofern diese Vorstellung nicht als auf das Verhältnis zu einer potenziell veränderbaren Wirklichkeit begrenzt beschrieben wird. Andererseits wird wiederum nicht gesagt, dass wir in diesem Verhalten berechtigt *sind*, sondern dass wir uns für berechtigt *halten*. Dies entkräftigt zwar die erste Aussage nicht, aber relativiert sie in unklarer Weise, insofern eine Wahrheit, die „ein für allemal" gültig ist, tatsächlich zu dem Verhalten berechtigt und deshalb diese näherliegende Formulierung erwartbar gewesen wäre. Ungeachtet der Frage, was ein „wahrer Geist" genau sei, ist zudem unklar, wie James zur Ansicht gelangt, dass – auch wenn nur in rhetorischem Sinne – ein Prozent solcher Vorgehensweisen problematisch wäre. Kann es sein, dass die vieldiskutierte Skepsis gegenüber solchen Induktionsschlüssen bei James auf diese kleine Randnotiz geschrumpft ist? James' abschliessender Satz ist zudem, so meine ich, ohne Berücksichtigung seines Wahrheitsbegriffs überhaupt nicht zu verstehen. Die Betrachtung eines Exemplars einer Gattung führt zum wahren Urteil über die gesamte Gattung, das sich dadurch als Wahrheit erweise, weil es nicht widerlegt wird? Man könnte hier zuerst an Popper denken, der selbstverständlich einverstanden wäre, etwas als wahr zu bezeichnen, das nie falsifiziert wird – nur wird dieser Zeitpunkt bekanntlich nie eintreten. Das heisst wiederum auch, dass die Wahrheit dieses Urteils nie bewiesen wäre. Durch die ausreichende Diskussion des pragmatistischen Wahrheitsverständnisses lässt sich jedoch erahnen, was James hier meinen *muss*: dass es sich hier um eine Methode handelt, die zu einem wahren Urteil führt, insofern kein Widerspruch bekannt *ist*, und solange kein Widerspruch erfahren *wird*. Blendet man James' erwähnte Wahrheit über die Welt aus, die seine eigenen Vorstellungen von Wahrheit torpediert, ist jedenfalls zu erkennen, dass seine Variante des Pragmatismus enumerative Induktion „nach Gattungen" zulässt. Jedenfalls wäre es hier auch die einzige Stelle, an der James effektiv Kritik an aktuellen wissenschaftlichen Vorgehensweisen üben würde, wäre diese Schlussart ihm zufolge nicht zulässig.

9. Ein Versuch der pragmatistischen Rechtfertigung induktiver Schlüsse

Gehen wir nun mit Fokus auf die Induktion über James' Ausführungen hinaus, stellt sich eine erste Frage. Wenn wir die Minimaldefinition für induktive Schlüsse beiziehen, die darin besteht, dass aufgrund von Beobachtungen auf Unbeobachtetes geschlossen wird, scheint ein Vorteil des pragmatistischen Wahrheitsverständnisses auf dem Spiel zu stehen: dass Wahrheit durch Erfahrung erzeugt wird. Die Lösung des Problems um die Anforderung an wahre Vorstellungen, mit der Wirklichkeit übereinzustimmen, basiert darauf, dass Übereinstimmung erst durch die Interaktion mit der Wirklichkeit entsteht, da sie ansonsten leer ist und bestenfalls durch Zufall zustande kommt. Man könnte hier gerade diese Auffassung hinzunehmen, um gegen die Berechtigung induktiver Schlüsse zu argumentieren. Andererseits kann man argumentieren, dass durch die Projektion von Beobachtetem auf Unbeobachtetes diese Übereinstimmung nach pragmatistischem Verständnis gegeben ist – für die Prämissen – und schliesslich ebenso für die Konklusion, da darin ja die Eigenart solcher Schlüsse besteht. Insofern *aufgrund* von Beobachtetem geschlossen wird, wird das Verhältnis von Vorstellung und Wirklichkeit, wie es durch Beobachtung tatsächlich erzeugt worden ist, für Unbeobachtetes im gleichen Verhältnis vorausgesetzt. Das heisst im Umkehrschluss, dass nicht auf beliebiges Unbeobachtetes geschlossen werden darf, von dem gar nicht angenommen wird, dass es sich im gleichen Verhältnis zur Wirklichkeit verhalte, wie das ihm zugrundeliegende Beobachtete. In diesem Sinne kann von zufälliger Übereinstimmung gar nicht die Rede sein. Das klingt zweifellos trivial, da dies für gewöhnlich gar nicht zur Debatte steht und es beispielsweise bei der enumerativen Induktion als selbstverständlich gilt, ein Verhältnis zwischen Gegenstand und Eigenschaft als starre Beziehung auf unbeobachtete Gegenstände derselben Art zu übertragen. In diesem Sinne kann man argumentieren, dass verhindert wird, dass zwischen der erschlossenen Vorstellung und der Wirklichkeit die kritisierte Leere entsteht. Dennoch ist die Übereinstimmung zwischen der Vorstellung des Unbeobachteten und der Wirklichkeit von anderer Qualität als die zwischen der Vorstellung des Beobachteten und der Wirklichkeit. Das ist, wie es bereits bei der Begriffsdiskussion hervorgehoben wurde, generell eine nicht zu verhindernde Eigenschaft induktiver Schlüsse. Wenn wir die obige Interpretation der Ausführungen James' hinzuziehen, wonach gewisse induktive Urteile wahr sind, wenn sie keine Widerlegung erfahren – einerseits da zum Urteilszeitpunkt keine dem Urteil widersprechende Erfahrung vorliegt (Kohärenz) und andererseits, weil das Urteil so lange gilt, bis eine widersprechende Erfahrung gemacht wird – können wir mit dem Pragmatismus noch weiter gehen: Wir können den induktiven Schluss als Ganzes als mit der Wirklichkeit übereinstimmend betrachten. Dass alles Brot mich nähren könnte, stimmt demnach mit der aktuell erfahrbaren Wirklichkeit überein. Wir leben in diesem Sinne in einer Wirklichkeit, in der gewisse Gesetze wahr sind und es ist alternativlos, diese anzunehmen, bis sie allenfalls falsifiziert sind – was unter Umständen noch immer auch „verfeinert" bedeuten kann – und unsere Wirklichkeit verändert wird. Bis dahin können wir zwischen den Gesetzen und der Wirklichkeit Übereinstimmung und folglich die Wahrheit der Gesetze erzeugen. Daran, dass wir dann Wissen über wahre Gesetze haben, ändert die mögliche Existenz einer nicht erfahrbaren objektiven Realität nichts. Ich denke auf diese Weise könnte man für die Wahrheit bestehender induktiv erschlossener Gesetze und Aussagen argumentieren: dass sie durch ihr Basieren auf der Erfahrung hinreichend mit der Wirklichkeit übereinstimmen und mit dieser und bestehenden Wahrheiten kohärent übereinstimmen müssen. Damit liesse sich auch die naiv anmutende Äusserung James' legitimieren, wonach es Gattungen anstatt lauter Einzelexemplare gebe. Obgleich er eigentlich nicht sagen kann, dass unsere Welt diese Eigenschaft „ein für allemal" hat, ist es doch so, dass die Existenz von Gattungen unserer Wirklichkeit entspricht und die blosse Denkbarkeit einer künftigen Erfahrung, die diese Wirklichkeit verändern würde, noch kein Grund ist, die Wahrheit, dass es Gattungen gibt, zu verwerfen. Jede denkbare Wahrheit, die mit Erfahrung zusammenhängt, hat meines Erachtens die Eigenschaft, dass sie auch als falsch denkbar ist. Alles, was jedoch bis zu ihrer tatsächlichen Falsifikation vorliegt, ist ihre Verifikation und Anwendbarkeit, die mit ihrer Übereinstimmung mit unserer Wirklichkeit einhergeht.

Dass mit diesem Wahrheitsverständnis für die Wahrheit induktiv erschlossener Gesetze und Aussagen argumentiert werden kann, scheint mir plausibel. Eine andere Frage, die sich wohl bei jeder Herangehensweise an das Induktionsproblem aufdrängt, ist diejenige nach den Bedingungen, unter denen auf solche Aussagen geschlossen wird oder werden darf. Wann offenbart sich ein Gesetz zum ersten Mal als unsere Wirklichkeit? Denken wir wiederum an die logische Form der enumerativen Induktion, so ist klar, dass es unzählige

Möglichkeiten gibt, ein induktives Argument zu bilden, dass niemand für sinnvoll erachten würde. Goodman führt hierzu die Frage an, ob die Bekanntschaft mit einem Mann, der sich gerade im Zimmer befindet und der dritte Sohn der Familie ist, die Hypothese stützt, dass alle Männer in diesem Zimmer dritte Söhne sind (1988, 97). Ungeachtet dessen, dass ein solcher Schluss bereits aufgrund des analytischen Gehalts seines Prädikats verworfen werden muss, weil es ohne erste und zweite keine dritten Söhne gibt, widerspricht er unserer Wirklichkeit. Wir haben auch Erfahrungen von ersten und zweiten Söhnen, wie auch davon, dass Aufenthaltsorte und familiäre Beziehungen zumeist in zufälligem Verhältnis stehen, und – falls dies hier nicht der Fall sein sollte – wissen meistens, ob wir in eine Zusammenkunft dritter Söhne geraten sind. Jedenfalls scheint klar, dass jemand, der aufgrund seines Zusammentreffens mit einem Mann, der ein dritter Sohn ist, einen solchen Schluss zieht, für in seinem Verhalten unberechtigt gehalten würde. Eine verlockende Bezeichnung für induktive Schlüsse, die für gewöhnlich akzeptiert werden, ist „die beste Erklärung". Harman und Lipton, die diesen Begriff in die Debatte um die Induktion eingeführt und für seine Verwendung argumentiert haben, sollen hier aussen vor gelassen werden. Ihre Formel lässt sich allerdings auch passend in die vorliegende Diskussion einfügen, da sie so verstanden werden kann, dass sie immer über die blossen Prädikate eines Induktionsschlusses hinausdeutet, was ein zentrales Merkmal für solche Schlüsse ist, die mehrheitlich akzeptiert werden. Die Frage, was eine gute Erklärung für das Auftreten einer Reihe von Erfahrungen ist, verweist immer auf andere, bereits anerkannte Wahrheiten und ruft somit die Anforderung der Kohärenz mit bestehendem Wissen auf den Plan. Nur wenn neue Erfahrungen mit sämtlichen bestehenden Wahrheiten, die durch sie nicht widerlegt werden, kohärent sind, können Erklärungen für Erfahrungen als mit der Wirklichkeit – wenn auch nur der Form nach – übereinstimmend gehalten werden. Dabei dürfte es der Fall sein, dass eine solche Erklärung genau dann für die bessere gehalten wird, wenn sie auch mehr inhaltliche Kohärenz mit der Wirklichkeit aufweist, weil ihre Begründung ebenso besser ist, wenn sie klarer und weitläufiger auf den Zusammenhang mit bisherigen Erfahrungen verweist. Wenn wir erklären, warum wir bislang nur grüne Smaragde gefunden haben und dies dadurch tun, dass wir schliessen, dass alle Smaragde grün seien, so ist dies besser begründet, wenn wir etwa auf die Farbkonsistenz anderer Minerale verweisen und den Zusammenhang von Farbkonsistenz und chemischer Beschaffenheit von Stoffen hinzuziehen. Sollten diese hier als Begründungen angeführten Wahrheiten umgekehrt erst durch den Schluss auf die Eigenschaft aller Smaragde erzeugt werden, ist auch besser ersichtlich, was James mit Übereinstimmung als „Vorgang des Geführt-Werdens" oder der „wirksamen Berührung mit der Wirklichkeit" zu meinen scheint: eine fortlaufende Vernetzung von Wahrheiten, die zu neuen Erfahrungen führen und umgekehrt. Es ist in diesem Prozess jedoch immer ausschlaggebend, dass wir die Wirklichkeit und somit die aktuellen Wahrheiten berücksichtigen und nicht nur monoton auf die Erfahrungsreihe als Prämissen eines induktiven Arguments, dessen logische Form ohnehin für nichts garantiert, schauen und hoffen, dass sich spontan eine Berechtigung des Schlusses ergibt.

Entsprechend ist auch zu Goodmans neuem Rätsel der Induktion zu sagen, dass die Möglichkeit einer solchen Eigenschaft, die er mit „grot" im Sinn hat, bloss denkbar, aber nicht Teil unserer Wirklichkeit ist. Der fundamentale Unterschied liegt darin, dass niemand darauf schliessen würde, dass alle Smaragde grün sind, wenn bereits Erfahrungen von sich verändernden Farben von Mineralien vorliegen. Insofern derartige Eigenschaften aber nicht Teil der Wirklichkeit sind, wäre es beliebig, eine nie erfahrene und neuartige Eigenschaft in den Induktionsschluss aufzunehmen. Wenn auf Basis von nie erfahrenen und bloss denkbaren Gegenständen geschlossen wird, ist es auch unmöglich, dem pragmatistischen Wahrheitsverständnis gerecht zu werden, denn falls sich der Schluss einst doch als verifizierbar herausstellen sollte, wäre dies komplett zufällig.

Ich denke, dass man ein pragmatistisches Verständnis von induktiven Schlüssen, angelehnt an James' Anforderungen an eine Theorie, folgendermassen skizzieren könnte:

Induktive Schlüsse sind gerechtfertigt, wenn sie die beste Erklärung für neue Phänomene sind. Eine Erklärung muss dabei mit möglichst vielen bisherigen Wahrheiten, denen die Phänomene nicht widersprechen, übereinstimmen, was als formale Kohärenz betrachtet werden kann. Dabei ist eine Erklärung desto besser, je mehr bestehende Wahrheiten sie stützen, beziehungsweise je mehr bestehende Wahrheiten durch die Erklärung gestützt werden, was inhaltliche Kohärenz oder Kohäsion genannt werden kann. In anderen Worten könnte man sagen, dass ein Induktionsschluss besser ist, je stärker er mit der Wirklichkeit interagiert und zu bestehenden Wahrheiten führt beziehungsweise je stärker bestehende Wahrheiten zu ihm hinführen. Es ist anzunehmen, dass die nach diesen Kriterien ausgewählte Erklärung stets auch empirisch überprüfbare Konsequenzen aufweist, die zumindest neue Erfahrun-

gen verheissen. Die am stärksten in dieser Art mit der Wirklichkeit vernetzte Erklärung kann als Wahrheit angenommen werden, weil es keine andere Erklärung gibt, die in höherem Masse mit der Wirklichkeit übereinstimmen würde. Dabei ist Wahrheit selbstverständlich wiederum pragmatistisch, als relativ zur aktuell erfahrbaren Wirklichkeit zu betrachten, sodass diese an epistemischen Grenzen orientierte Rechtfertigung für induktive Schlüsse nicht dadurch kritisiert werden kann, dass sich ihre Konklusion dereinst als falsch herausstellen könnte. Es ist hier jedoch nicht gesagt, dass Hypothesen nach der Form eines Induktionsschlusses keinen wissenschaftlichen Nutzen haben können, nur weil sie diesen Anforderungen nicht entsprechen. Mir scheint es schwierig abzuschätzen, ob Demokrits Atomismus – gesetztenfalls er sei wissenschaftshistorisch tatsächlich als Initialzündung der späteren Verifikationen dieser Theorie zu betrachten – in der geforderten Weise mit der damaligen Wirklichkeit in Übereinstimmung war. Andererseits ist es wahrscheinlich aber auch so, dass seine Hypothese auch nur als solche und nicht etwa als Wahrheit betrachtet wurde.

10. Abschliessende Kritik an der Kritik

Die Kritik, die meistens an der Praxis induktiven Schliessens geübt wird oder zumindest Ausgangspunkt der ihr entgegengebrachten Skepsis darstellt, ist meines Erachtens oft das Resultat von zu hohen Ansprüchen. Einer davon könnte darin liegen, dass, wie schon eingangs diskutiert, eine logische Wahrheitskonservierung nach dem Vorbild eines deduktiven Schlusses zur Norm erklärt und zur Bewertung der Berechtigung induktiver Schlüsse hinzugezogen wird. Ein anderer zu hoher Anspruch würde darin bestehen, dass die Gewissheit, die wir durch Beobachtung von Gegenständen erlangen, gleichermassen in Aussagen über Unbeobachtetes gesucht wird. Beides zeugt höchstens davon, was man sich von einer wissenschaftlichen Methode erhofft und ist offensichtlich unrealistisch. Was das dieser Untersuchung zugrundeliegende Wahrheitsverständnis zudem hat andeuten lassen, ist, dass scheinbar oftmals blosse Möglichkeiten und denkbare Ereignisse, die zu keinem Zeitpunkt Teil der Wirklichkeit waren, angeführt werden, um die Versuche, sich neue Erfahrungen durch bestehendes Wissen erklärbar zu machen, abzuqualifizieren. Meines Erachtens ist es jedoch mindestens so abwegig, reine Denkbarkeiten mit konkreten Erfahrungen gleichzusetzen, wie von Aussagen über Unbeobachtetes dieselbe Qualität zu erwarten, wie sie Aussagen über Beobachtetes aufweisen. Von Unbeobachtetem ist nichts aus besseren Gründen zu erwarten als das, was wir aufgrund von früheren Erfahrungen beziehungsweise der gegebenen Wirklichkeit davon erwarten.

Wenn Konklusionen induktiver Schlüsse im oben ausgeführten Sinn mit unserer Wirklichkeit übereinstimmen können – was sie durch höhere inhaltliche Kohärenz umso mehr tun – können sie auch als wahr erachtet werden, denn ein Wahrheitsbegriff, dessen Bedingungen ausserhalb epistemischer Grenzen liegen, ist schlicht nicht praktikabel. Doch auch wer es nicht für gerechtfertigt hält, induktiv auf solche pragmatistischen Wahrheiten zu schliessen, muss letztlich wohl doch zugeben, dass gewisse induktive Schlüsse alternativlos sind.

Literatur

- Broad, Charlie D. 1926. *The philosophy of Francis Bacon. An address delivered at Cambridge on the occasion of the Bacon tercentenary.* Cambridge: University Press.
- Da Costa, Newton C. A. and Steven French. 1989. "Pragmatic Truth and the Logic of Induction." *British Journal for the Philosophy of Science* (3) 40: 333–356.
- Fahrbach, Ludwig. 2016. "Scientific revolutions and the explosion of scientific evidence." *Synthese. An International Journal for Epistemology, Methodology and Philosophy of Science* (12) 194: 5039–5072.
- Goodman Nelson. 1988. *Tatsache, Fiktion, Voraussage.* Übersetzt von Philippi, Bernd und Hermann Vetter. Frankfurt a. M.: Suhrkamp.
- Graeser, Andreas. 2000. *Bedeutung, Wert, Wirklichkeit. Positionen und Probleme: Texte zur Philosophie des 20. Jahrhunderts.* Bern: Lang.
- Hume, David. (1748) 1964. *Eine Untersuchung über den menschlichen Verstand*, herausgegeben von Manfred Kühn, übersetzt von Raoul Richter. Hamburg: Meiner.
- James, William. 1908. *Der Pragmatismus: Ein neuer Name für alte Denkmethoden. Volkstümliche philosophische Vorlesungen.* Übersetzt von Wilhelm Jerusalem. Leipzig: Dr. Werner Klinkhardt.
- Musgrave, Alan. 2004. "How Popper [Might Have] Solved the Problem of Induction." *Philosophy* (79) 307: 19–31.
- Popper, Karl R. (1935) 1994. *Die Logik der Forschung.* Die Einheit der Gesellschaftswissenschaften, Bd. 4, 10., verb. und verm. Aufl., Tübingen: J.B.C. Mohr Paul Siebeck.
- ———. 1973. *Objektive Erkenntnis. Ein evolutionärer Entwurf.* Hamburg: Hoffmann u. Campe.
- Scheller, Jörg. 2020. „Die Virologie ist derzeit ein ähnliches Stadtgespräch wie die Relativitätstheorie in den Zwanzigern." *Neue Zürcher Zeitung*, 24. April 2020. https://www.nzz.ch/feuilleton/die-virologie-ist-derzeit-ein-aehnliches-stadtgespraech-wie-die-relativitaetstheorie-in-den-zwanzigern-ld.1552961
- Will, Ulrich. 1981. *Das Problem der Induktion.* Diss. Phil. Köln: Print.

Micha Herrmann (27) befindet sich am Ende seines Masterstudiums in Germanistik und Philosophie. Der pragmatistische Wahrheitsbegriff nach James hat ihn nach anfänglicher Ablehnung wiederholt beschäftigt und schliesslich zum Versuch motiviert, diesen für eines seiner Interessengebiete, die Wissenschaftsphilosophie, nutzbar zu machen.

A Defense of Perspectivism about Ought Against the Argument from Advice

1. Introduction: Objectivism and Perspectivism

In moral philosophy, there is a debate about whether what one ought to do depends on one's epistemic perspective. Perspectivists assume that it does, while objectivists deny this and claim that what one ought to do is always determined by all facts. According to Benjamin Kiesewetter (2011), the two views can be characterized as follows.

Objectivism
S ought to φ if, and only if, *S* ought to φ relative to all facts (known and unknown).

Perspectivism
S ought to φ if, and only if, *S* ought to φ relative to the facts, that are within *S*'s epistemic perspective.

Naturally, the question then arises as to what is meant by an epistemic perspective. Different versions of perspectivism have different understandings about what constitutes the epistemic perspective of a person. For example, some assume that a person's actual beliefs constitute their epistemic perspective. Others take the view that the evidence available to a person constitutes their epistemic perspective (Kiesewetter 2011, 3). This essay will not deal with a particular version of perspectivism, since my current aim is to defend perspectivism against an objection that is not directed against a particular perspectivist view but criticizes the basic assumption that is common to all perspectivist accounts.

The disagreement between objectivists and perspectivists is well illustrated in the following example offered by Frank Jackson.

> **Doctor I**
> Jill, a doctor, has to decide on the correct treatment for her patient, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the available evidence has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of drugs B and C will completely cure the skin condition; the other though will kill the patient, and there is no way that she can tell which of the two is the perfect cure and which the killer drug. In fact, drug B is the cure, while C will lead to the patient's death. What ought Jill to do? (Jackson 1991, 462-463)

To answer this question, according to objectivism, one must consider all the facts. In Jill's situation, the fact that the patient would only be completely cured by taking B is a decisive reason for Jill to give B to the patient. Jill, therefore, ought to give B to the patient, since this is what she ought to do relative to all facts. The perspectivists' answer is different. They assume that in answering the question of what Jill ought to do, one only has to consider the facts that are within Jill's epistemic perspective. Since Jill cannot possibly know which drug will completely cure the patient and which will kill him, the fact that the patient would be completely cured by B has no bearing on what Jill ought to do. From her epistemic perspective, she ought to give A to the patient, because in trying to guess the cure, there is a 50% risk of killing the patient, which is not outweighed by the 50% chance of curing his minor skin condition (Kiesewetter 2011, 5).[1]

At this point, one could think that the disagreement between objectivists and perspectivists can be resolved by

1 Examples like Doctor I are sometimes used to argue against objectivism. It is argued that objectivism gives the wrong answer to the question, "What ought Jill to do?" (Kiesewetter 2011, 5–6). The question of whether an argument against objectivism can be derived from examples like Doctor I is relevant to the debate but is not the subject of the present essay.

distinguishing between different notions of "ought." The idea here is that there is both an objective and a subjective notion of "ought." The objective notion of "ought" refers to what a person ought to do relative to all the facts. The subjective notion of "ought" refers to what a person ought to do relative to an epistemic perspective. In Doctor I, for example, Jill ought to *objectively* (i.e., relative to all facts) give treatment B, but *subjectively* (i.e., relative to her perspective) give treatment A. This suggests that objectivists and perspectivists are talking about different things. While objectivists tell us what Jill ought to do objectively, perspectivists tell us what Jill ought to do subjectively. Therefore, the distinction between an objective and a subjective notion of "ought" seems to reveal that there is no substantial disagreement between objectivists and perspectivists (Kiesewetter 2011, 2).

However, Kiesewetter has argued that the distinction between different notions of "ought" does not show that there is no substantial disagreement between objectivists and perspectivists. When Jill deliberates on what to do, she seems to ask only one question, namely, "What ought I to do?" Therefore, whether the "ought" in Jill's question depends on all the facts or her perspective seems to be a substantive question that is answered differently by the two views. In the following, I will therefore assume that there is a substantial disagreement between objectivists and perspectivists, which consists in the fact that they have different views on which notion of "ought" we use in the deliberative question, "What ought I to do?" (Kiesewetter 2011, 2).

My current goal is to defend perspectivism against a common objection. The objection states that perspectivism provides an account of our practice of advice that is implausible. I call this objection *The Argument from Advice*. In section II, I will reconstruct this argument. In section III, I will first discuss a response to the argument from advice which was advanced by Elinor Mason (2013). She argues for an alternative perspectivist account of our practice of advice. While her account can avoid some of the problems arising from the argument from advice, I will show that it ultimately fails to provide a convincing response to it. In section IV, I will argue for a third perspectivist account that avoids all the problems arising from the argument from advice as well as the problems that Mason's account faces. In section V, I will compare this third account to the objectivist account of our practice of advice. Although the objectivist account fits our linguistic intuitions better, I will argue that the perspectivist approach is more convincing because it provides a more plausible account of our practice of advice.

2. The Argument from Advice

One of the most influential objections against perspectivism is what I will call *The Argument from Advice*. The basic idea of this argument is that perspectivism involves an implausible conception of our practice of advice. The argument begins by applying perspectivism to said practice. It is then shown that the resulting perspectivist conception of this practice faces three crucial problems. In the reconstruction of this argument, I will show that we can derive a requirement for a plausible account of our practice of advice from each of these problems. The formulation of these requirements is important both because they help to understand the basic idea of the argument from advice, and, more importantly, because they serve to evaluate the different accounts of our practice of advice that are discussed in this essay.

How perspectivism is applied to our practice of advice is well illustrated by the following example, which is a modification of Doctor I.

> **Doctor II**
> Jill is in the same situation as in Doctor I. Jill is about to give A to the patient when Jack, a colleague of Jill's, enters the room. Since Jill does not know which of the two drugs B and C would cure the patient completely and which would kill the patient, she asks Jack, "What ought I to do?" Jack knows that drug B would cure the patient completely and drug C would kill the patient. So, he advises Jill: "You ought to give B to the patient." (Kiesewetter 2011, 7)

As we have already seen in the example of Doctor I, objectivists and perspectivists understand the term "ought" in different ways. This also applies to the use of "ought" in Jill and Jack's respective statements. Now the question arises as to how perspectivism understands Jill's question and Jack's advice.

Let's start with Jill's question. According to perspectivism, Jill's question refers to what she ought to do from her epistemic perspective. Thus, when she asks, "What ought I to do?" she is actually asking, "What ought I to do from my epistemic perspective?" Benjamin Kiesewetter has argued that this interpretation of the question seems to be wrong. "When Jill asks the adviser, 'What ought I to do?' it seems that she does not want a report about what her current evidence already tells her to do, but rather hopes for something

beyond that" (Kiesewetter 2015, 7). If we interpret the "ought" in Jill's question as a subjective "ought" that is relative to Jill's current epistemic perspective, then we cannot account for the natural thought that Jill hopes to get additional information about the situation that could help her in determining what she ought to do. By generalizing this observation about Jill's question in Doctor II, we can formulate the first requirement for a plausible account of our practice of advice.

> **Requirement 1 (R1)**
> A plausible account of our practice of advice must account for the natural idea that a person who asks an adviser what they ought to do seeks information that goes beyond their current epistemic perspective.

What about the perspectivist interpretation of Jack's advice? According to perspectivism, Jack makes a statement about what Jill ought to do *from her epistemic perspective*. However, this seems to be wrong. Instead, his advice seems to focus on what she objectively ought to do, namely to give the patient treatment B. Judith Thomson formulates the basic idea of the problem as follows: "On those rare occasions someone conceives of asking my advice on a moral matter, I do not take my field work to be limited to a study of what he believes is the case: I take it incumbent upon me to find out what is the case" (Thomson 1986, 179). When we give a person advice about what to do in a specific situation, we do not seem to be limited by that person's current perspective. Therefore, the second requirement is as follows:

> **Requirement 2 (R2)**
> A plausible account of our practice of advice needs to account for the natural idea that an advisor who advises a person *P* about what they ought to do, is not limited by *P*'s current epistemic perspective.

Another problem with the perspectivist interpretation is that, according to it, Jack's advice is literally false. This is because, as claimed by perspectivism, Jill ought to give the patient A and not B. However, this judgment contradicts the natural idea that Jack, precisely because he is better informed, gives Jill the only correct advice in this situation. Thus, according to perspectivism, it is impossible for an adviser to give correct advice about what a person ought to do if her perspective suggests doing something else. This also applies when the adviser is better informed than the advisee, which is implausible (Kiesewetter 2015, 7). Therefore, the third requirement is as follows:

> **Requirement 3 (R3)**
> A plausible account of our practice of advice needs to account for the natural idea that an adviser who advises a person *P* to φ, while *P*'s epistemic perspective does not suggest φ-ing, can do so correctly.

In summary, the argument from advice is directed against the perspectivist account of our practice of advice. I will call this account the *Subjective Ought Account*.

> **Subjective Ought Account (SOA)**
> (i) A person who asks an adviser what they ought to do, is asking about what they ought to do relative to their epistemic perspective.
> (ii) An adviser who tells a person *P* that they ought to φ, is saying that *P* ought to φ relative to *P*'s epistemic perspective.

The argument from advice states that perspectivism does not provide a plausible account of our practice of advice because SOA does not meet R1–R3. It thus represents a serious objection to perspectivism.

3. Mason's Perspectivist Account

In her essay "Objectivism and Prospectivism about Rightness", Elinor Mason (2013) tries to defend perspectivism against the argument from advice. In doing so, she argues for an alternative perspectivist account of our practice of advice that avoids two of the problems that SOA faces. In the following section, I will first reconstruct Mason's account and show how it meets R2 and R3. I will then argue that her account cannot refute the argument from advice, firstly because it does not meet R1, and secondly because it gives rise to a new problem that SOA does not face.

In her defense of perspectivism against the argument from advice, Mason argues that we have to interpret the "ought" in Jack's advice differently than SOA does. According to her, Jack's advice does not refer to what Jill ought to do from *Jill's* epistemic perspective, but to what Jill ought to do from *Jack's* epistemic perspective. Therefore, when Jack says, "You ought to give B to the patient," he actually means, "From my epistemic perspective, you ought to give B to the patient" (Mason 2013, 11). It is important to emphasize that Ma-

son's interpretation of Jacks advice is *perspectivist*. This is because Jack uses a subjective notion of "ought" in his advice that is relative to *his* epistemic perspective. Since Mason only changes the second part (ii) of SOA, her account is as follows:

Mason's Perspectivist Account (MPA)

(i) A person who asks an adviser what they ought to do, is asking about what they ought to do relative to their epistemic perspective.

(ii) An adviser *A* who tells a person *P* that they ought to φ, is saying that *P* ought to φ relative to *A*'s epistemic perspective.

In order to evaluate the plausibility of MPA and to determine whether it is more convincing than SOA, we will now look at whether MPA meets R1–R3. I will do this by using the example of Doctor II. MPA fulfills R2 since according to MPA Jack's advice is not limited by Jill's epistemic perspective. Jack is only limited by his own perspective. This allows him to go beyond Jill's epistemic perspective when advising Jill. MPA also meets R3, as Jack's advice is true according to this account. Jack's advice refers to what Jill ought to do relative to *his* epistemic perspective. From Jack's perspective, Jill ought to give B to the patient because he knows that drug B is the perfect cure. The crucial point here is that both Jill's judgment that she ought to give A to the patient and Jack's advice that Jill ought to give B to the patient are true. This is so because they have different senses of "ought" in mind. While Jill uses a subjective notion of "ought" that is relative to *her* epistemic perspective, Jack uses a subjective notion of "ought" that is relative to *his* epistemic perspective (Kiesewetter 2015, 8). Thus, although according to perspectivism a person *P* ought to do what *P* ought to do relative to *P*'s epistemic perspective, the advice of an adviser *A* can still be true, since it refers to what *P* ought to do relative to *A*'s epistemic perspective.

An obvious problem with Mason's account is that it keeps the first part (i) of SOA, which concerns Jill's question. Therefore, the problem that arises from SOA's interpretation of Jill's question affects MPA as well. MPA cannot account for the natural idea that Jill seeks information that goes beyond her current epistemic perspective. From this, it follows that MPA does not satisfy R1.

MPA's interpretation of Jack's advice is also problematic. This is because according to MPA, Jill and Jack do not talk about the same subject matter. As we have already seen, Jill's question refers to *what Jill ought to do from her perspective*. Jack's advice, on the other hand, refers to *what Jill ought to do from his perspective*. As a result, they talk past each other. Jack's advice, therefore, does not actually answer Jill's question. It follows from MPA that we are forced to give up the intuitive idea that giving advice is about answering a person's question "What ought I to do?" for him or her (Kiesewetter 2015, 8). These considerations suggest the formulation of a further requirement.

Requirement 4 (R4)

A plausible account of our practice of advice must take into account the natural idea that a person *P* who asks an adviser *A* what they ought to do and *A* who tells *P* that *P* ought to φ use the same notion of "ought."

In summary, it can be said that MPA satisfies R2 and R3 but is still not convincing because it does not meet R1 and R4. MPA, therefore, does not provide a convincing reply to the argument from advice. Thus, perspectivists have good reasons to look for an alternative account of our practice of advice that can circumvent the problems raised by the argument from advice. Such an account is discussed in the following section.

4. The Conditional Subjective Ought Account

In this section, I will first present a third perspectivist account of our practice of advice, which is different from both SOA and MPA. I will call this account *The Conditional Subjective Ought Account* (CSOA). This account can be found in Kristian Olsen (2017). Afterward, I will argue for the plausibility of CSOA by showing that CSOA meets R1–R4. As we have already seen with MPA, it is not enough to just interpret the adviser's advice differently than SOA does, because otherwise R1 cannot be met. CSOA, therefore, reinterprets both the adviser's advice and the advisee's question. To illustrate this, CSOA will be explained using the example of Doctor II.

Let's start with Jill's question. According to CSOA, what Jill actually means when she asks, "What ought I to do?" is this: "What ought I to do if I had your (Jack's) epistemic perspective?" (Olsen 2017, 369). What is important here is that Jill does not ask what she ought to do from Jack's epistemic perspective. That would be a problem for CSOA because intuitively, Jill is asking about what she ought to do from her own perspective, not what she ought to do from Jack's perspective (Kie-

sewetter 2015, 7). However, CSOA can account for this intuition because according to CSOA, Jill asks about what she ought to do from her own epistemic perspective *if it were a different one*, namely Jack's.

How does CSOA interpret Jack's advice? What Jack means when he tells Jill, "You ought to give A to the patient," is this: "If you had my (Jack's) epistemic perspective, you ought to give A to the patient." Thus, Jack is also concerned with what Jill ought to do from her epistemic perspective if she had his (Jack's) epistemic perspective. Therefore, according to CSOA, both Jill and Jack use the term "ought" in a *subjective* and *conditional* sense. The "ought" they refer to is subjective because it is relative to Jill's epistemic perspective and conditional because it is about what Jill ought to do if she had a different (in this case, Jack's) epistemic perspective (Olsen 2017, 369). We are now able to formulate CSOA.

Conditional Subjective Ought Account (CSOA)

(i) A person *P* who asks an adviser *A* what they ought to do, asks about what they ought to do if they had *A*'s epistemic perspective.

(ii) An adviser *A* who tells a person *P* that they ought to φ is saying that *P* ought to φ if *P* had *A*'s epistemic perspective.

Let us now look at Doctor II to see if CSOA meets R1–R4. CSOA satisfies R1 because, according to CSOA, Jill does not want to know what she ought to do from the epistemic perspective that she currently has, but what she ought to do from her perspective if it were the same as Jack's. Thus, according to CSOA, Jill seeks information that goes beyond her *current* epistemic perspective. CSOA meets R2 because, according to CSOA, when Jack tells Jill to give B to the patient, he is not limited by Jill's current epistemic perspective. However, according to CSOA, Jack's advice is plausibly limited by *his* epistemic perspective as this would be the perspective of Jill if she had the same one as Jack. According to CSOA, Jack's advice is: "If you had my epistemic perspective, you ought to give B to the patient." This conditional sentence is true, although Jill ought to give A to the patient from her current epistemic perspective. CSOA therefore meets R3. And finally, CSOA also meets R4 because, according to CSOA, Jack and Jill are talking about the same thing, namely what Jill ought to do if she had Jack's epistemic perspective. Thus, both use the same notion of "ought" in their respective statements, which is both subjective and conditional. This prevents Jill and Jack from talking past each other on the one hand, and on the other hand, it accounts for the intuition that Jack's advice answers Jill's question for her.

In summary, CSOA can avoid the three problems arising from SOA, since CSOA meets R1–R3. In addition, compared to MPA, CSOA represents a more plausible perspectivist account of our practice of advice because CSOA fulfills R4 and thus circumvents the problem that has arisen from MPA. I therefore conclude that endorsing CSOA seems to be a good strategy for perspectivists to respond to the argument from advice.

5. CSOA and the Objectivist Account

At this point, the question may arise as to whether CSOA fits our linguistic intuitions in cases like Doctor II. Is it plausible to interpret the dialogue between Jill and Jack the way CSOA does? Would the straightforward objectivist reading not be preferable to this unintuitive interpretation? These are the questions that will be addressed in this section. I will begin by introducing the objectivist account of our practice of advice. We will see that this account, while meeting R1–R4, faces a serious problem. Finally, I will argue for the plausibility of CSOA by showing that CSOA is not affected by the problem that the objectivist account faces.

The objectivist account of our practice of advice states that adviser and advisee both talk about an objective notion of "ought." The objective "ought" is relative to all the facts and is therefore independent of any epistemic perspective. The account can be formulated as follows.

Objective Ought Account (OOA)

(i) A person *P* who asks an adviser *A* what they ought to do, asks about what they ought to do relative to all the facts.

(ii) An adviser *A* who tells a person *P* that they ought to φ is saying that *P* ought to φ relative to all the facts.

Let us now look at the example of Doctor II to see if OOA meets R1–R4. According to OOA, Jill asks about what she ought to do *independently of her epistemic perspective*. OOA thus takes into account the natural idea that Jill seeks information that goes beyond her current epistemic perspective. The objectivist account, therefore, meets R1. The objectivist account states that Jack's advice refers to what Jill ought to do independently of her epistemic perspective and thus accounts for the natural idea that Jack's advice is not limited by Jill's epistemic perspective. In his answer to Jill's question, he goes beyond Jill's epistemic perspective because he knows some relevant facts that she does not. Therefore,

OOA meets R2. According to OOA Jack's statement that Jill ought to objectively give B to the patient is true since Jill ought to give B to the patient relative to all facts. OOA therefore meets R3. OOA also states that Jill and Jack are both using the objective notion of "ought" in their respective statements. Both are therefore talking about the same subject matter, namely, *what Jill ought to do objectively*. Hence, OOA also meets R4.

In summary, OOA, like CSOA, meets the requirements for a plausible account of our practice of advice, which were formulated in the course of this essay. However, OOA has a decisive advantage over CSOA: it fits our linguistic intuitions much better. OOA thus seems to be the most plausible account of our practice of advice. If this turns out to be true, we have a good reason to reject perspectivism. However, I do not think that OOA is more plausible than CSOA. To see why, consider the following example by Olsen (2017), which is a modification of Doctor I.[2]

> **Doctor III**
> Jill, a doctor, has to decide on the correct treatment for her patient, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the available evidence has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of drugs B and C will completely cure the skin condition; the other though will kill the patient. Before deciding which drug to give to the patient she goes to Jack for advice about what to do. She knows that Jack has better evidence than she does but that he does not know which of the drugs is the perfect cure. She asks Jack, "What ought I to do in my circumstances?" Jack knows that the partial cure is C (not A, as Jill believes), but he's unsure which of A or B is the perfect cure and which is the killer. In light of this, Jack tells Jill, "You ought to give drug C."[3] (Olsen 2017, 368)

Do Jill and Jack talk about the objective "ought"? This does not seem to be the case. If the "ought" in Jill's question is interpreted objectively, it has the implausible consequence that Jill asks Jack what she ought to do relative to all the facts, knowing at the same time that Jack does not know all the facts. Hence an objectivist interpretation of Jill's question in Doctor III seems implausible.

Olson (2017) notices another problem that arises when we interpret the "ought" in Jack's advice objectively. If the "ought" in Jack's advice is interpreted objectively, it has the implausible consequence that we are forced to say that Jack's advice is literally wrong. However, this judgment contradicts the natural idea that Jack, precisely because he is better informed, correctly advises Jill (Olson 2017, 368). Thus OOA in Doctor III is confronted with the same problem as SOA in Doctor II. It cannot account for the natural idea that an adviser who bases his advice on better information gives correct advice. In general, the problem of OOA is therefore that it cannot adequately explain situations such as Doctor III, where a better-informed adviser *A* tells a person *P* what *P* ought to do, while *A* does not know all the relevant facts.

Let us take a look at how CSOA deals with Doctor III. According to CSOA, Jill and Jack both talk about a conditional subjective "ought." Jill's question poses no problems for CSOA. Jill wants to know what she ought to do if she had the same epistemic perspective as Jack. Even if she knows that Jack does not know which is the perfect cure, it makes sense for her to ask Jack for advice about what she ought to do since he might know more than she does. CSOA can also account for the intuition that Jack's advice seems to be correct. His advice is correct because it is true that Jill ought to give C if she had Jack's epistemic perspective. CSOA thus accounts for the natural idea that Jack advises Jill correctly about what she ought to do.

Examples like Doctor III present a challenge for objectivists because OOA cannot account for these examples. Because of this difficulty, we have a good reason to favor CSOA over OOA, even though OOA fits our linguistic intuitions better.

6. Conclusion

In this essay, I have tried to defend perspectivism about ought against the argument of advice. This argument shows that the most obvious perspectivist account of our practice of advice, SOA, is implausible and challenges perspectivists to provide an alternative account of our practice of advice. Elinor Mason (2013) has proposed such an alternative account. I have argued that her account is not convincing because, according to it, the adviser and the advisee do not talk about the same

2 Olsen (2017) presents this example as a modified version of a case by Kolodny and MacFarlane (2010, 121).

3 The editors pointed out that the examples Doctor II and III might be problematic from a gender perspective because the woman (Jill) always asks the man (Jack) for advice. However, I did not change the examples because I did not formulate them myself.

thing and consequently talk past each other. I then argued for CSOA, a perspectivist account that makes the following two claims: First, when we ask an adviser what we ought to do, we ask about what we ought to do if we had their epistemic perspective. Second, when we tell a person that they ought to φ, we are saying that they ought to φ if they had our epistemic perspective. It turned out that this account can avoid the problems of SOA and MPA. Finally, I defended CSOA against the objection that it does not fit our linguistic intuitions by showing that OOA, the account of our practice of advice that best fits our linguistic intuitions, is not convincing. This suggests that CSOA is superior to the other accounts of our practice of advice discussed in this essay, namely SOA, MPA, and OOA. If CSOA turns out to be the best account of our practice of advice, then—contrary to the widespread belief that our practice of advice poses a challenge to perspectivism—it provides a good reason to favor perspectivism over objectivism. While CSOA provides perspectivists with a plausible account of our practice of advice, objectivists face the challenge of finding one. Of course, I have not proven that CSOA provides the best account of our practice of advice. What I have shown, however, is that CSOA not only provides perspectivists with a good response to the argument of advice, but also challenges objectivists to develop a more plausible account of our practice of advice.

References

- Jackson, Frank. 1991. "Decision-theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101 (3): 461–482.
- Kiesewetter, Benjamin. 2011. "'Ought' and the Perspective of the Agent." *Journal of Ethics and Social Philosophy* 5 (3): 1–24.
- Kolodny, Niko, and John MacFarlane. 2010. "Ifs and Oughts." *Journal of Philosophy* 107 (3): 115–143.
- Mason, Elinor. 2013. "Objectivism and Prospectivism about Rightness." *Journal of Ethics and Social Philosophy* 7 (2): 1–21.
- Olsen, Kristian. 2017. "A Defense of the Objective/Subjective Moral Ought Distinction." *The Journal of Ethics* 21 (4): 351–373.
- Thomson, Judith Jarvis. 1986. "Imposing Risks." In *Rights, Restitution, and Risk: Essays in Moral Theory*, edited by William Parent, 173–191. Cambridge, MA: Harvard University Press.

David Lussi (22) studiert Philosophie und Erziehungswissenschaft im Master. Er interessiert sich allgemein für Probleme der Metaethik und besonders für Fragen, die mit Gründen und Normativität zu tun haben, sowohl im praktischen als auch im epistemischen Bereich.

Overdemandingness and Supererogation

Are Theories that Demand to Supererogate Necessarily Overdemanding?

1. Introduction

In 1958, J. O. Urmson tried to reshape the landscape of the deontic statuses actions may have. He argued that there are not only obligatory, forbidden, and merely permissible actions, but also—in his words—"saintly" or "heroic" ones (1958, 199). Many philosophers have found it appealing that there might be a further category of actions; actions that are—according to a first approximation—good to do but not bad not to do (see Chisolm 1963; Feinberg 1968; Richards 1971). Ever since, philosophers have been puzzled by what is now referred to as "supererogation" and how to properly define it.

Some years after Urmson's influential publication, in 1972, the utilitarian Peter Singer famously tried to show that people should do much more for worse off fellow human beings than most of them actually do. For some, Singer's reasoning provided a reason to donate more money or to change their lifestyles. According to others, it highlighted how demanding utilitarianism as a moral theory actually is; complying with its principles is very costly for moral agents and might demand too much from them.

The current philosophical debate about so-called overdemandingness objections in ethics is shaped by the idea that these two discussions are connected in a certain way. According to many philosophers working in the field (e.g., Wolf 1982; Hooker 2000; Vessel 2010; Chappell 2020) and what I will call *the implication thesis*, supererogation and overdemandingness are intertwined in the following way: if a moral theory denotes a supererogatory action as obligatory, then that moral theory is overdemanding. The goal of this paper is to challenge said view. I will argue there is no convincing pair of accounts of "overdemandingness" and "supererogation" for which the implication thesis is true. That conclusion is reached after introducing supererogation in section 2, formally elaborating and presenting various accounts of what it could mean for a moral theory to be overdemanding in section 3, and after pairing the most convincing ones of these accounts with different conceptions of supererogation in order to test the implication thesis in section 4. The last section concludes my reasoning and raises remaining questions.

2. Supererogation and the *implication thesis*

Perhaps unsurprisingly, philosophers are not in agreement on how to define "supererogation". For that reason, the concept is best introduced by means of paradigmatic cases:

(1) Heroism: An out-of-control trolley hurtles towards five construction workers. As the brave woman she is, Anna jumps on the tracks and stops the trolly—but her grit costs Anna one of her legs. Had she not intervened, the five workers would all have suffered severe injuries.

(2) Lifestyle: Beth is well aware of global inequality and climate change. For that reason, she changes her lifestyle drastically: she starts living in a tent, sells almost all her stuff and donates every penny she does not need to survive.

(3) Generosity: By chance, Colin becomes involved in a conversation with a stranger. The stranger, for some reason, tells him that she is reasonably wealthy and that she has a penchant for good wine. Out of generosity, Colin walks over to the nearby store and buys an expensive wine for his new acquaintance.

What unites these cases is that, intuitively, Anna, Beth, and Colin all do something good which they would not have to do. They do more than what is actually demanded from them.

A lot has been written about whether various moral theories are compatible with supererogation and to what extent this is a problem for them (see for example Williams 1973; Scheffler 1982; Guevara 1999; Murphy 2000; Greene 2013, and for various defences, see Kagan 1989; Singer 1993; Unger 1996). It is widely believed that without profound modulation neither consequentialism, Kantianism, nor moral pluralism (see Ross [1930] 2002) leave room for supererogatory actions.

For my purposes, however, it is of no particular relevance whether certain famous moral theories leave room for supererogation. What gets the discussion going instead are two jointly sufficient observations: (i) there are supererogatory actions,[1] and (ii) for every supererogatory action, at least one moral theory denoting that action as obligatory can be constructed.[2]

If one then proceeds to look at theories according to which Anna and Beth, for instance, in fact act as they *should*, one will probably come to think that obeying these theories is very hard for moral agents. According to such theories, agents (at least sometimes) have to resist the urge to protect their bodies from harm or they are not allowed to live their lives as they initially have planed. They are not allowed to have even a somewhat expensive hobby, they may not give presents to their friends, and they should not sleep in a heated room during winter (at least not under current circumstances, that is). Such theories, one might say, are asking too much from moral agents; obeying their principles is too costly.

In view of such considerations, it might seem plausible that demands to supererogate (or, more precisely, demands to do what in fact is supererogatory)[3] imply overdemandingness. Compare, for instance, Susan Wolf who writes: "[i]f, as I have argued, this means that we have reason to want people to live lives that are not morally perfect, then any plausible moral theory must make use of some conception of supererogation" (1982, 438). According to Wolf, it seems, any moral theory which does not leave any room for supererogation demands too much. What is as well indicative of the connection between the supererogatory and overdemandingness is the fact that both Wolf and Urmson (1958) write about moral saints and saintly actions—the former in the context of overdemanding moral theories, the latter when describing supererogation.

Other philosophers seem to accept the connection as well. Vessel uses the expressions "Too Demandingness objection" and "Lack of Supererogation objection" against utilitarianism interchangeably (2010, 300), Chappell maintains that "by requiring actions that are intuitively supererogatory, maximizers are subject to the objection that their conception of morality is overly demanding" (2020, 500), and Hooker notes that "[a]ct-consequentialism is normally taken to be unreasonably demanding, construing as duties what one would have thought were supererogatory self-sacrifices" (2000, 149).

To spell out what I call the *implication thesis* (henceforth IT) more clearly, one might say that these philosophers take the following inference to be conclusive:

(1) φ-ing is supererogatory
(2) According to theory T, φ-ing is obligatory

---

(3) ∴ T is overdemanding

As the inference is not deductively valid the way it stands, the question whether it is conclusive depends on what is meant with the concepts used. In the subsequent sections I will discuss different accounts of overdemandingness and supererogation and look at the implications they have for the plausibility of IT.

Before doing so, two preliminary remarks seem appropriate. One is concerned with the compatibility of overdemandingness objections and the belief that morality *should* demand quite a lot from us, the other one is about the fact that negating IT does not mean to deny that there is any connection between supererogation and overdemandingness.

In my view, it is perfectly reasonable to be sceptical about overdemandingness objections in ethics. As Raz notes, claims that moral theories are too demanding are "liable to seem suspect" (Raz 1993, 1297), since we typically do not want people to shirk their moral duties by simply stating that they feel to have done enough already. In light of global inequality, climate change,

1 It should, however, not be concealed that there are philosophers who reject that there are genuine supererogatory actions. See, for instance, Moore ([1903] 1994) or Pybus (1982).

2 Here is a short guideline: take action φ which is supererogatory for agent A and define moral theory T in such a way that it follows from T that A ought to φ (e.g., by making "A ought to φ" a basic moral principle of T).

3 Strictly speaking, statements like: "A ought to φ and φ-ing is supererogatory for A" seem to be conceptually contradictory. For simplicity, however, "to demand to supererogate" is used synonymous with "to demand to do something that is in fact supererogatory" throughout this paper.

or animal suffering, morality *should* be challenging, it *should* demand a lot from us, and it seems that our moral duties reach beyond what most of us are currently doing. Additionally, as Rachels mentions, we do not want moral philosophers to become "orthodoxy's most sophisticated defenders, assuming that the existing social consensus must be right, and articulating its theoretical 'justification'" (1991, 70). All that being said, there still seem to be situations in which what would be the best thing to do is not demanded of moral agents. There does not appear to be a moral duty to give presents to (reasonably wealthy) strangers even if doing so might be better (and maybe praiseworthy) compared to not doing so, for instance. Additionally, it seems desirable that moral theories are able to incorporate the intuition that what Anna and Beth are doing is somehow *heroic*. If they would simply act as they ought, their actions would not be as praiseworthy as we typically take them to be.

For those reasons, overdemandingness poses an important philosophical problem that has to be taken seriously, even (or maybe especially) by those who want morality to be demanding. I for my part will take it to be one of the main goals for people working in this field that they are able to bring together the strong and heavy moral demands of a globalized, occasionally dangerous, and oftentimes unjust world with the existence of overdemanding moral theories and supererogation.

The second preliminary remark is that even though I deny IT, I nonetheless do not claim that there is no connection between supererogation and overdemandingness at all. In fact, I agree with the idea that by modifying a moral theory in such a way that it then comprises supererogation, that theory will *probably, ceteris paribus*, be less demanding.[4] This is so because chances are quite high that according to a moral theory which leaves no room for supererogation, once room for supererogation is made, some of the newly supererogatory actions will have been demanded beforehand. If this is so, then there is again some chance that those demands that were dissolved were (at least partly) responsible for making the theory overdemanding. Therefore, if your moral theory leaves no room for supererogation and is overdemanding, it might be a good idea to try to make room for it, even if IT eventually turns out to be false.

4 This idea can be found in many places. Ashford (2003, 282–283), for instance, seems to implicitly accept it when discussing the demands of Scanlon's (1998) contractualism.

3. Five conceptions of overdemandingness

Raising overdemandingness objections against moral theories is much easier than explicating what it means for a moral theory to be overdemanding. Let me therefore start my examination of IT by examining different conceptions of "overdemandingness".

3.1 Overdemandingness in the broad sense

Firstly, overdemandingness can be understood in a broad sense where it refers to extensional adequacy. Let me explain. Most moral theories attribute deontic statuses to actions: they tell us for instance that we are obligated to save a drowning child (see Singer 1972) or that it is forbidden to push the *fat man* in front of the trolly (see Thomson 1985).[5] At the same time, it is evident that most theories sometimes attribute wrong moral statuses, for example if they always, *all things considered*, take it to be forbidden to lie to Nazis or if they demand from us that we run into a burning building to save a cat. In the former case too much is forbidden, in the latter too much is demanded. "Too much" hereby means that there is at least one action that belongs to the set of actions that are demanded by a given theory but not to the set of actions that are actually demanded. The set of demands made by such a theory therefore reaches beyond the set of actual moral demands, making it overdemanding, so to speak.

It needs to be noted that a theory needs not demand much from us to be overdemanding in the broad sense. Take, for instance, a theory which merely demands that people have to light a candle on the seventh of December in order to celebrate Noam Chomsky's birthday. Such a theory demands very little from us, but since there does not seem to be an actual, universal duty to celebrate Chomsky's birthday, such a theory would still be overdemanding in the broad sense. Additionally, one can observe that many famous moral theories seem to be overdemanding in the broad sense. Kantianism and act utilitarianism, for instance, might be overdemanding in this sense as the examples of the Nazis and the burning building respectively seem to show.[6] One might argue, however, that the claim that a moral theory is overdemanding in the broad sense is in itself not the strongest of objections one can make towards it. This

5 Note, however, that there are exceptions. Scalar utilitarianism, for instance, does not attribute any deontic statuses to actions. Rather, it merely tells us what is morally better or worse (see Norcross 2006; Tobia 2017).

6 Kantianism might be argued to be over*demanding* in the sense that the negative duty not to lie seems to imply the demand to always say the truth—which, in turn, does not seem to be a member of the set of all actual moral demands.

is because such objections do not, for instance, unveil an inconsistency in the theory itself and because they can always be addressed by means of a bite-the-bullet reply (which, admittedly, might not be equally acceptable in all cases).

Returning to the implication thesis, it is evident that IT is almost trivially true if overdemandingness is understood in the broad sense: once a moral theory denotes a supererogatory action as obligatory, its set of demanded actions exceeds the set of actual moral demands at one point, making the theory overdemanding. It seems to be equally evident, however, that people maintaining IT have something other than overdemandingness in the broad sense in mind when linking supererogation with overdemandingness. Compare, for instance, Singer's claim that "if it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, to do it" (1972, 231). People rejecting the so-called "Singer principle" (Arneson 2004, 33) due to its demandingness typically will not simply claim that it is not actually demanded to always prevent bad things from happening if doing so is possible without sacrificing anything of comparable moral importance. Rather, those people seem to suggest that there is a further, deeper problem in Singer's utilitarian moral theory. Singer, they would say, overestimates how hard, challenging, demanding, or costly it may be to comply with a moral theory.

3.2 Overdemandingness in the pure sense

This leads us to what McElwee calls "the pure demandingness objection" (2017, 89) where costs for the agent obtain a more central role. According to McElwee, objection O against moral theory T (according to which, due to some consideration C, agent A has a duty to φ) is a pure demandingness objection *iff* O claims (i) that C is insufficient to generate a duty to φ, and (ii) that this is the case precisely because φ-ing would be too costly for A. Condition (ii) is needed in order to distinguish the pure overdemandingness objections from other criticisms, namely the "wrong moral ranking objection" and the "wrong overall ranking objection". The former claims that the reason why C is insufficient to generate a duty to φ is that C is outweighed by other moral considerations. The latter, on the other hand, claims that C is outweighed by non-moral considerations (McElwee 2017, 89–90).

It needs to be noted that the pure overdemandingness objection has an important thing in common with overdemandingness in the broad sense: it is concerned with extensional inadequacy as well. The central difference is that instead of merely claiming a certain extensional inadequacy, the pure overdemandingness objection provides a reason for why the proclaimed inadequacy occurs (which is what condition (ii) is here for). Pure overdemandingness objections therefore always conceptionally entail an objection that the theory at hand is extensionally inadequate.

As it stands now, McElwee's notion of "overdemandingness" cannot yet be contrasted to overdemandingness in the broad sense, since that concept was applied to moral theories themselves whilst McElwee is interested in characterising *objections* to moral theories. What needs to be done, therefore, is to ascertain which conception of "overdemandingness" the pure overdemandingness objection could be based on. A major clue lies in the observation that the pure overdemandingness objection, due to condition (ii), seems to assume that there is a certain threshold τ beyond which complying with a moral demand becomes too costly (and therefore making the theory extensionally inadequate), no matter which other moral or non-moral reasons are at hand. In other words: there is not even one moral demand for which it is true that complying with it would lead to costs higher than τ.

Thus, I take the pure overdemandingness objection to correspond to the following conception of "overdemandingness":

> Moral theory T is overdemanding in the pure sense *iff* there exists at least one moral demand D to engage in action φ deducible from T for which it is true that there is at least one agent A who is an addressee of D and for whom the costs of φ-ing would exceed threshold τ.

Given the definition of τ given above, that account of "overdemandingness" entails extensional inadequacy.

Whether there actually are overdemanding moral theories in the pure sense consequently depends on whether threshold τ exists, and people might be hesitant to accept τ. In addition, it should be noted that defining τ might be even more difficult than one might think on first sight. One challenge might be that it seems to be both reasonable and difficult at the same time to differentiate between the amount of cost that may result from duties which can be fulfilled in one go on the one hand (e.g., saving a drowning child) and from duties which influence an agent's whole life (e.g., donating a certain share of one's monthly income). It seems to me that people accepting τ should also accept that costs can be higher if they are distributed over time (although, admittedly, this depends on how costs are defined). Defenders of

overdemandingness in the pure sense should be able to include this consideration when defining τ, for instance by defining different thresholds ($\tau_1, \tau_2, \ldots, \tau_n$) where the subscripts correspond to relevantly different types of moral demands.[7] As I shall demonstrate in turn, however, there are further conceptions of overdemandingness that do not rely on some threshold τ.

3.3 Overdemandingness as wrong moral ranking

As mentioned in the previous section, McElwee discusses two further kinds of overdemandingness objections which are relevant for my purposes here: the wrong moral ranking objection and the wrong overall ranking objection. Analogously to what had to be done in the case of the pure overdemandingness objection, these two objections need to be transformed into accounts of "overdemandingness" as well. As a preliminary note, it needs to be mentioned that just like the pure overdemandingness objection, these two objections entail extensional inadequacy as well. The difference between the three objections lies in how the inadequacy is said to come about.

The wrong *moral* ranking objection opposes that a moral theory proclaims a moral duty where there are stronger moral reasons in favour of a non-compatible alternative. At the same time, it is claimed that whenever there are stronger moral reasons for non-φ than for φ, there cannot be a duty to φ. Consequently, we reach the following conception of "overdemandingness":

> Moral theory T is overdemanding in the moral ranking sense *iff* there exists at least one moral demand D to engage in action φ deducible from T for which it is true that there is at least one agent A who is an addressee of D and for whom it is true[8] that they have stronger moral reasons for non-φ than for φ.

Could this be an adequate schematisation of what people making overdemandingness objections have in mind? I take it that most people would agree that if a moral theory claims that there is a duty to engage in a certain action whilst there seem to be stronger moral reasons counting against engaging in that action, then that moral theory is extensionally inadequate. What I am doubting, however, is that people really take a moral theory to make a moral ranking mistake when criticising it as overdemanding. If one objects to Kantianism that it disregards further relevant moral considerations when claiming that one has to tell the truth to Nazis looking for persecuted people, then one typically would not say that the claim is that Kantianism is overdemanding. Similarly, if utilitarians demand that we push the *fat man* in front of the trolley, we would say that they fail to ascribe enough weight to other relevant moral reasons. Nevertheless, we would not say that, because it follows from it that we should push the *fat man*, utilitarianism is overdemanding. "Overdemandingness" therefore does not seem to be concerned with moral ranking.

3.4 Overdemandingness as wrong overall ranking

The wrong *overall* ranking objection conversely corresponds to a more plausible understanding of what it means to be overdemanding. Instead of claiming that a moral theory makes a mistake when weighing up reasons within the moral domain, it maintains that a theory comes to wrong conclusions when balancing moral against non-moral considerations. Using the same line of reasoning as in the case of the wrong moral ranking objection, we reach the following conception of "overdemandingness":

> Moral theory T is overdemanding in the overall ranking sense *iff* there exists at least one moral demand D to engage in action φ deducible from T for which it is true that there is at least one agent A who is an addressee of D and for whom it is true (i) that the moral reasons they have for φ weigh *more* than the moral reasons they have for non-φ and (ii) that the total of all moral and non-moral reasons they have for φ *weighs* less than the total of all the moral and non-moral reasons they have for non-φ.[9]

This account of "overdemandingness" bears a lot of resemblance to the one discussed beforehand. Nonetheless, people who understand "overdemandingness" in the overall ranking sense and who at the same time maintain that there in fact are overdemanding moral theories commit themselves to a strong claim about practical reasons which defenders of the previous account do not have to

7 If this were to be done, the definition of overdemandingness in the pure sense given above would have to be modified slightly: one would have to replace ("threshold τ" with "the for D relevant threshold τ_x").

8 Note that the formulation "for whom it is true" does not refer to some sort of epistemic relativism or reason internalism. Rather, the formulation is derived from the fact that moral reasons are always reasons *for* someone.

9 The first condition is necessary in order to guarantee that when looking solely at moral reasons, φ-ing would come out on top. If this were not the case, T would already be overdemanding in the *moral* ranking sense. Also, note that the two conditions jointly imply that there are more non-moral reasons for non-φ than for φ.

make. Namely, they have to claim that it is possible for non-moral reasons to outweigh moral ones. This commitment entails a second conjecture, namely that it is even conceptually possible (and, ideally, also epistemically feasible[10]) to weigh up moral and non-moral reasons.

3.5 Overdemandingness as confinement

In 2016, Benn added a further conception of "overdemandingness" into the mix which does not rely on any of McElwee's overdemandingness objections. According to Benn, a moral theory can also be overdemanding if it is overly-confining.

Benn starts her reasoning by examining what she calls the "Tripartite View" (2016, 77), which is the same position that Urmson tried to refute in 1958 (see introduction), namely that there are only three deontic statuses: the obligatory, the morally indifferent, and the forbidden. Her objection to this view is that it claims that every action that is not morally indifferent is either demanded or forbidden and that because of this, moral duties are "ubiquitous". The follow-up question, Benn asks and answers herself, is in what sense moral theories, where duties are ubiquitous, are overdemanding. The response she gives is that, in her own words, "[a]ny theory that is overly-confining is overly-demanding, not because of the content of what it demands, but because it overly limits what we can permissibly do" (Benn 2016, 78). Consequently, Benn's account of "overdemandingness" seems to look as follows:

> Moral theory T is overdemanding in the confinement sense *iff* T is overly-confining, which is the case if T denies supererogation (e.g., because T entails the Tripartite View[11]).

If "overdemandingness" is understood in such a way, it might at first seem as if IT was true. Of course, this has to do with the fact that developing an account of "overdemandingness" which entails a conceptual connection to supererogation was Benn's goal all along. Nonetheless, I do not think that Benn's work can be taken to show that IT is true. One of two reasons for this is that Benn's notion of "overly-confining" does not seem to capture what people have in mind when they object to moral theories that they are too demanding. Granted, an overly-confining moral theory puts moral agents in an unpleasant position that is neither desirable nor extensionally adequate. However, if people object to moral theories, like act utilitarianism, that they are overly demanding, their claims are directed to the content of the moral principles deductible from these theories. Someone who maintains that a theory telling us that we ought to do what Anna, Beth, and Colin (see section 2) are doing is overdemanding, clearly thinks that the fault lies in these demands themselves, rather than in further assumptions about deontic statuses the theory makes. Benn's account of "overdemandingness" does not seem to be able to do justice to this. That being said, I maintain that Benn is right in claiming that being overly-confining is a severe problem for a moral theory. However, I do not think that people maintaining IT would want to subscribe to Benn's way of defining "overdemandingness".

The second reason is grounded in the fact that there are two ways to think about whether "overdemandingness" and "supererogation" are conceptionally connected. The first one is the way of IT, namely to think about whether demands to carry out individual supererogatory actions make a moral theory overdemanding. The second way conversely takes a more general approach. It asks whether moral theories leaving no room for supererogation in general are overdemanding—and this seems to be what Benn is concerned with. Obviously, this is philosophically interesting in its own right, but it will not be relevant for deciding about the conclusiveness of IT. One might object that the second way entails the first one since theories neglecting supererogation in general necessarily have to demand actions which are supererogatory. However, this is not true. What would be needed is the additional premise that a theory represents what would be supererogatory as demanded instead of denoting it as morally indifferent or forbidden. As soon as this premise is added, the discussion becomes relevant for the conclusiveness of IT again, but only at the cost of focussing on individual supererogatory actions again.

4. Supererogation and the *implication thesis* – part II

Having presented all these different views of what it could mean for a moral theory to be overdemanding, I will now look at which accounts of supererogation these views would have to be paired with in order to make IT true. Before doing so, let me summarise which accounts are already eliminated from the outset. *Overdemandingness in the broad sense* was deemed irrelevant for IT since it renders IT to be trivially true and because it cannot account for the observation that overdemandingness objections are not merely about extensional inadequacy.

10 The claim in brackets is not entailed in the first commitment.

11 Theories entailing what one might call the Bipartite View (i.e., the view that there are only demanded and forbidden actions), as utilitarianism is sometimes thought to be doing, deny supererogation as well.

Overdemandingness in the moral ranking sense faced a similar problem: overdemandingness does not seem to be concerned with *moral* ranking. Benn's *overdemandingness as confinement* was deemed irrelevant for IT as well due to the fact that IT is concerned with the demandingness of specific moral demands, rather than with the demandingness of general assumptions about deontic statuses a moral theory makes. Having said that, we are left with *overdemandingness in the pure sense* and *overdemandingness in the overall ranking sense* for the remaining part of the paper.

4.1 Supererogation and costs for the agent

Overdemandingness in the pure sense, as I have shown in section 3.2, ascribes high importance to costs for the agent. Thus, it seems like IT is true if the right account of supererogation entails costs for the agent as a necessary condition as well. The idea behind this reasoning would be that if only costly actions are supererogatory and if "overdemandingness" would be understood in the pure sense, then every moral theory that demands to supererogate would pose too high costs on us, making it overdemanding.

The goal of the current section is to show that this reasoning is flawed for two reasons. Firstly, I claim that it is unreasonable to commit oneself to the claim that supererogatory actions are necessarily costly. And secondly, I argue that even if that commitment was plausible, it might not be enough to guarantee IT's truth.

There is a huge body of literature on whether supererogatory actions always involve significant costs for the agent. Many philosophers maintaining what Benn in a later paper calls the "extreme cost view" (2018, 2406) claim that they do (e.g., Rawls 1971; McGoldrick 1984; Straurnanis 1984; Jackson 1986; Stanlick 1999)[12]. Rawls, for instance, claims that the high costs he takes all supererogatory actions to bring about are the very reason for why supererogatory actions are not demanded (1971, 117). If we look at those cases which defenders of supererogation typically invoke first, this view may appear plausible. Cases like Anna's or Beth's are usually designed precisely in such a way that they yield high costs for their protagonists. Once we look at cases like *Generosity*, however, it seems that proponents of the extreme cost view have difficulties to explain what is going on. Regarding Colin, it does not seem to be true that he faces high costs due to his actions. Additionally, it does not seem to be the case that the reason why Colin is not obligated to buy wine lies in the costs it would have for him. Rather, it seems that there simply is no universal duty to give presents to reasonably wealthy strangers.

At this point, proponents of the extreme cost view may choose one of two replies. Either they can choose to argue that what Colin is doing is not supererogatory at all but rather demanded (or—and this might be even more implausible—morally indifferent or even forbidden), or they can try to introduce a further deontic status. That status would then, just like supererogation, be exemplified by actions that are (i) better than their omissions and (ii) not demanded (i.e., their omissions—although being worse—are allowed). The additional condition of being costly would then, so the strategy, be reserved for supererogation alone. The first strategy is highly implausible. Claiming that Colin does as he should or that he does not do anything morally significant at all, simply does not describe the situation in a satisfactory way. I furthermore suppose that the second strategy is undesirable as well. Reserving the predicate of being supererogatory for a certain subset of actions for which conditions (i) and (ii) from above hold true (namely for the ones that are costly) does not do justice to what the basic idea of supererogation seems to be and to the reason why it was introduced by Urmson in the first place. Urmson's observation was simply that there are actions which are good to do on the one hand but are not demanded on the other hand. High costs for the agent are not relevant at all in this first approximation of the concept. Also, to use a different approximation, actions which *go beyond the call of duty* (which is where the expression of "supererogation" etymologically originates from) do not seem to be costly by conceptual necessity.

As announced previously, I think that defenders of IT (who understand "overdemandingness" in the pure sense) would still face a significant difficulty, even if supererogatory actions were necessarily costly. That difficulty is to show that the amount of cost that is needed in order to make an action supererogatory is the same amount of cost that threshold τ requires in order to render a moral demand to be overly demanding. Prima facie, there do not seem to be strong reasons to think that these two costs should necessarily be equally high—unless one adopts what may be called the "Rawlsian approach" to supererogation, according to which the reason why supererogatory actions are not demanded is that they are so costly. The Rawlsian, therefore, could claim that the amount of cost that is needed in order to convert a demanded action into a supererogatory one is precisely the same amount of cost that makes complying with a moral demand too costly (i.e., the costs go beyond threshold τ). The downside of this account is, of course, that it faces

12 For the opposing standpoint, confer Archer (2016) or Benn (2018).

the very same objections that were raised one paragraph ago: for some supererogatory actions, the reason for why they are not demanded seems to be distinct from the costs these actions would bring to the agent. Also, some presumably supererogatory actions are not costly at all. The Rawlsian cannot account for these observations.

Without a Rawlsian approach, however, as just mentioned, there is no particular reason to think that φ-ing should become supererogatory precisely when a moral demand to φ would become overly demanding. It would resemble an inexplicable miracle if these two predicates would be co-extensional without them being connected conceptually in a Rawlsian sense.

Additionally, given that sometimes moral demands may be quite difficult and costly to adhere to, defenders of IT would be at risk of shrinking the extension of supererogatory actions much more than would be plausible. If there is a duty for privileged people to donate quite a lot of money, to go vegan, not to fly on vacation, or to host refugees, then—if supererogation is necessarily costly—actions need to be much more costly in order to have a chance of being supererogatory than seems to be plausible.

To summarise; *overdemandingness in the pure sense* is of little help when defending IT. Even pairing it with the—as I have argued to be implausible—claim that supererogatory actions are necessarily costly is not sufficient to demonstrate IT's truth. What would be needed is a Rawlsian account of supererogation. The problem with that account, however, besides being extensionally inadequate, is that the reason for the optionality of the supererogatory does not seem to be that these actions are too costly for being obligatory. What should not be overlooked either is the philosophical work defenders of IT still owe us regarding how to define and defend threshold τ or the set of thresholds $\{\tau_1, \tau_2, \dots, \tau_n\}$ discussed in subsection 3.2.

4.2 Supererogation, non-moral reasons and overridingness

This last subsection will begin with a brief excursion into the field of practical reasons. Subsequently, I proceed to show how defenders of IT might try using *overdemandingness in the overall ranking sense* to their advantage. I will defend their argument against a potential worry and raise a further objection.

While deliberating, moral agents often face a broad variety of different reasons. Some of those are moral ones (e.g., the fact that it saves their life is a moral reason to save a drowning child), others are of non-moral kind (e.g., the fact that it ruins my suit is a non-moral reason against jumping in the pond)[13]. One important and philosophically challenging question moral agents need to ask themselves at this stage is how moral and non-moral reasons may interact in deliberation processes. According to what has been called *overridingness* of moral reasons (see Portmore 2008), moral reasons trump non-moral ones. This means that whenever there are stronger moral reasons for action φ than for the non-compatible alternative ψ, then even the strongest non-moral reasons counting in favour of ψ-ing are overridden by the moral reasons counting in favour of φ-ing.[14] A major benefit people might see in *overridingness* is that it might appear to be a necessary condition for moral rationalism, i.e., the view that if an action is morally demanded, then it is what there is most overall reason to do. As Portmore (2008) and Archer (2014) have argued, however, proponents of moral rationalism need not rely on *overridingness*. If they are right, it is conceptually possible for non-moral reasons to outweigh moral ones. For people wishing for a demanding morality, this might—similar to overdemandingness objections—seem suspect. It appears that another backdoor has opened through which immoral persons might flee from the call of their moral duties. Note, however, that the mere possibility of outweighing moral with non-moral reasons does not entail that this can be done easily. It is reasonable to assume that against the normative force of moral reasons only the strongest of non-moral reasons stand a chance, even if *overridingness* is false.

Returning to the last remaining conception of overdemandingness, *overdemandingness in the overall ranking sense*, it appears that it presupposes that *overridingness* is false. Only if that is the case, conditions (i) and (ii) mentioned in subsection 3.4 can be fulfilled jointly. In order to find out the implications this has for IT, a deeper investigation of supererogation is needed.

Some action φ is supererogatory for agent A only if there is a non-compatible alternative, ψ , and if it is true (a) that A is both morally allowed to φ and to ψ, and (b) that A has more moral reason to φ than to ψ. Condition (b) is needed—and here I agree with Dorsey (2013, 359)—in order to show what is *super*-erogatory about φ-ing (or, in other words, why φ-ing is morally meritorious). Nevertheless, as Portmore discusses, supererogation

13 Arguably, this might only be an instrumental (but still non-moral) reason which helps to promote the agent's own interest. Many philosophers agree that agents have a non-moral rather than a moral reason to promote their own interest (see Bratman 1994; Portmore 2008).

14 The overridingness of moral reasons should not be confused with what Stroud (1998) refers to with "overridingness", namely a kind of moral rationalism.

might appear "almost paradoxical" (2008, 379) in light of these two conditions: how can ψ-ing be morally allowed given that there is a non-compatible alternative that has stronger moral reasons counting for it?

Defenders of IT (who reject *overridingness*, as just seen), can provide the following explanation: ψ-ing is morally permissible (and φ-ing is not morally demanded) because the moral reasons for φ-ing are outweighed by the non-moral reasons for ψ-ing (see Portmore 2008, 380). If one merely examines the moral reasons for actions φ and ψ, φ-ing comes out on top. Once the non-moral reasons for ψ are considered as well, one can notice that they may hinder the moral reasons for φ from generating a duty to φ. From there, the defenders may continue their plea for IT. Overdemanding moral theories, they might say, are the ones that demand to φ (for which there is more moral reason than for the non-compatible alternative ψ) even though, once both moral and non-moral reasons for both φ and ψ are considered, there is more reason to ψ. And since this depiction of φ, ψ and the reasons for those actions fits the picture of supererogation drawn just now, they might conclude that theories demanding to do what is supererogatory are overdemanding by conceptual necessity.

Before presenting my objection to that argument, I will defend it against another worry one might have. The worry is that, analogously to the first objection to Benn's account, *overdemandingness in the overall ranking sense* is not a plausible account of "overdemandingness" at all. Consequently, it might be the case that the concept comprises a conceptual connection to supererogation without in fact being the concept of "overdemandingness". The reason why I do not share this worry is that it seems to me that disallowing people to hold on to the non-moral reason they have to promote their own interest as soon as even the weakest of moral reasons for a non-compatible alternative appears, is highly demanding. It demands that people disregard their own interest to a much wider extend than can be demanded from them. Returning to Wolf, this seems to fit her picture of a rational moral saint quite well; the picture of a person who "pays little or no attention to his [sic!] own happiness in light of the overriding importance he [sic!] gives to the wider concerns of morality. […] [T]his person sacrifices his [sic!] own interests to the interests of others, and feels the sacrifice as such" (1982, 420). Demanding from someone to be a rational moral saint seems to be too high of a demand.

Let me now turn to what I find not to be convincing in the defence of IT stemming from *overdemandingness in the overall ranking sense*. My objection, in a nutshell, is that the argument presented above is not deductively valid. In order to see why, a closer look at all the reasons for φ and ψ is necessary. Let me name the moral reason for action φ "r1", the non-moral reason for φ "r2", the moral reason for action ψ "r3" and the non-moral reason for ψ "r4". According to *overdemandingness in the overall ranking sense*, a theory demanding to φ is overdemanding *iff* (i) r1 is the stronger reason than r3 and (ii) r1 and r2 together are weaker than the pair of r3 and r4. Regarding supererogation, φ-ing is only supererogatory if (iii) r1 is stronger than r3 and (iv) r4 is stronger than r1. What the defenders of IT needed to claim was that it is overdemanding to demand engaging in an action whose moral reason for doing it is outweighed by a non-moral reason for doing a non-compatible alternative (i.e., a supererogatory action). A closer look at conditions (i)–(iv) reveals, however, that φ-ing can be supererogatory even if it is not the case that r1 and r2 together are weaker than the pair of r3 and r4. This is so because condition (iv) only entails a certain relation between r4 and r1; r2 on the other hand can be arbitrarily strong. Consequently, not all instances of supererogation fulfil the condition that they render moral theories overdemanding if they regard them as morally demanded. This gap between "supererogation" and "overdemandingness" shows that there is no such conceptual connection as proclaimed by IT.

5. Conclusion

The philosophical debate about overdemandingness objections in ethics is shaped by the idea that failing to recognise supererogatory actions as what they are and instead denoting them as obligatory, renders moral theories to be overly demanding. Over the last few pages, I examined this view more closely and argued that it is not accurate. The first step towards that conclusion was to show that there are genuine supererogatory actions. After a preliminary comment about the relation between overdemandingness objections on the one hand and the strong moral demands of a globalised world on the other hand, I presented five accounts of what it could mean for a moral theory to be overdemanding. I argued that overdemandingness objections neither solely object to a theory's extension of moral demands nor to the ranking of moral reasons it makes. Additionally, I showed that what I called the *implication thesis* is concerned with the demandingness of individual moral demands, rather than with the demands that stem from deeper claims about the moral landscape made by a moral theory. Because of this, Benn's account of overdemandingness was deemed irrelevant for examining the implication thesis. In section 4, I paired the last two remaining conceptions

of "overdemandingness" with various views about supererogation and looked at whether there are constellations in which the implication thesis appears plausible. I argued that both *overdemandingness in the pure sense* and *overdemandingness in the overall ranking sense* are of no help in order to demonstrate IT's truth. Against the former, I objected that it either needs to be paired with the implausible Rawlsian approach to supererogation or that it cannot explain why the costs needed for supererogation are equally high as the ones required for surpassing threshold τ. To the latter, my response was that even if rejecting *overridingness* is plausible, there is conceptual room for cases of supererogation where the moral and non-moral reasons for that action jointly are not weaker than the combination of moral and non-moral reasons for a non-compatible alternative. This showed that not all demands to supererogate need to be overdemanding. What remains an unanswered question for now is how supererogation relates to various forms of moral rationalism. Additionally, if supererogation and overdemandingness are conceptually as different as I think they are, it will be interesting to see whether overdemandingness can—similarly to supererogation—pose new challenges in all kinds of areas of philosophical inquiry.

References

- Archer, Alfred. 2014. "Moral Rationalism without Overridingness." *Ratio* 27 (1): 100–114.
- ———. 2016. "Supererogation, Sacrifice, and the Limits of Duty." *The Southern Journal of Philosophy* 54 (3): 333–354.
- Arneson, Richard. 2004. "Moral Limits on the Demands of Beneficence?" In *The Ethics of Assistance: Morality, Affluence, and the Distant Needy*, edited by Deen K. Chatterjee, 33–58. Cambridge: Cambridge University Press.
- Ashford, Elizabeth. 2003. "The Demandingness of Scanlon's Contractualism." *Ethics* 113 (2): 273–302.
- Benn, Claire. 2016. "Over-Demandingness Objections and Supererogation." In *The Limits of Moral Obligation: Moral Demandingness and Ought Implies Can*, edited by Marcel van Ackeren and Michael Kühler, 68–83. New York: Routledge.
- ———. 2018. "Supererogation, optionality and cost." *Philosophical Studies* 175 (10): 2399–2417.
- Bratman, Michael E. 1994. "Kagan on 'the appeal of cost'." *Ethics* 104 (2): 325–332.
- Chappell, Richard. 2020. "Deontic Pluralism and the Right Amount of Good." In *The Oxford Handbook of Consequentialism*, edited by Douglas W. Portmore, 498–512. Oxford: Oxford University Press.
- Chisolm, Roderick. 1963. "Supererogation and Offence: A Conceptual Scheme for Ethics." *Ratio* 5 (1): 1–14.
- Dorsey, Dale. 2013. "The Supererogatory, and How to Accommodate It." *Utilitas* 25 (3): 355–382.
- Feinberg, Joel. 1960. "Supererogation and Rules." *Ethics* 71 (4): 276–288.
- Greene, Joshua. 2013. *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. London: Penguin Press.
- Guevara, Daniel. 1999. "The Impossibility of Supererogation in Kant's Moral Theory." *Philosophy and Phenomenological Research* 59 (3): 539–624.
- Hooker, Brad. 2000. *Ideal Code, Real World*. Oxford: Oxford University Press.
- Jackson, M. W. 1986. "The Nature of Supererogation." *The Journal of Value Inquiry* 20: 289–296.
- Kagan, Shelly. 1989. *The Limits of Morality*. Oxford: Oxford University Press.
- McElvee, Brian. 2017. "Demandingness Objections in Ethics." *The Philosophical Quarterly* 67 (266): 84–105.
- McGoldrick, Patricia M. 1984. "Saints and Heroes: A Plea for the Supererogatory." *Philosophy* 59 (230): 523–528.
- Moore, G. E. (1903) 1996. *Principia Ethica: Revised Edition*. Cambridge, MA: Cambridge University Press.
- Murphy, Liam. 2000. *Moral Demands in Nonideal Theory*. New York: Oxford University Press.
- Norcross, Alastair. 2006. "The Scalar Approach to Utilitarianism." In *The Blackwell Guide to Mills Utilitarianism*, edited by Henry R. West, 217–232. Oxford: Blackwell Publishing.
- Portmore, Douglas W. 2008. "Are Moral Reasons Morally Overriding?" *Ethical Theory and Moral Practice* 11 (4): 369–388.
- Pybus, Elizabeth M. 1982. "Saints and Heroes." *Philosophy* 57 (220): 193–199.
- Rachels, James. 1991. "When Philosophers Shoot from the Hip." *Bioethics* 5 (1): 67–71.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Raz, Joseph. 1993. "A Morality Fit for Humans." *Michigan Law Review* 91 (6): 1297–1314.
- Richards, David A. J. 1971. *A Theory of Reasons for Action*. Oxford: Clarendon Press.
- Ross, William David. (1930) 2002. *The Right and the Good*. Oxford: Clarendon Press.
- Scanlon, Thomas. 1998. *What we owe to each other*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Scheffler, Samuel. 1982. *The Rejection of Consequentialism*. Oxford: Clarendon Press.
- Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1 (3): 229–243.
- ———. 1993. *Practical Ethics*, Second Edition. Cambridge: Cambridge University Press.
- Stanlick, Nancy A. 1999. "The Nature and Value of Supererogatory Actions." *Journal of Social Philosophy* 30 (1): 209–222.
- Straurnanis, Joan. 1984. "Duties to Oneself: An Ethical Basis for Self-Liberation?" *Journal of Social Philosophy* 15 (2): 1–13.
- Stroud, Sarah. 1998. "Moral Overridingness and Moral Theory." *Pacific Philosophical Quarterly* 79 (2): 170–189.
- Thomson, Judith Jarvis. 1985. "The Trolley Problem." *The Yale Law Journal* 94 (6): 1395–1415.
- Tobia, Kevin. 2017. "A Defense of Scalar Utilitarianism." *American Philosophical Quarterly* 54 (3): 283–294.
- Unger, Peter. 1996. *Living High and Letting Die: Our Illusion of Innocence*. New York: Oxford University Press.
- Urmson, J. O. 1958. "Saints and Heroes." In *Essays in Moral Philosophy*, edited by A. I. Melden, 198–216. Seattle: University of Washington Press.
- Vessel, Jean-Paul. 2010. "Supererogation for Utilitarianism." *American Philosophical Quarterly* 47 (4): 299–319.
- Williams, Bernard. 1973. "A Critique of Utilitarianism." In *Utilitarianism: For and Against*, edited by J. J. C. Smart and Bernard Williams, 77–150. Cambridge: Cambridge University Press.
- Wolf, Susan. 1982. "Moral Saints." *The Journal of Philosophy* 79 (8): 419–439.

Timo Junger (23) schliesst im kommenden Sommer seinen Master in Philosophie ab. Er interessiert sich insbesondere für analytische praktische Philosophie und beschäftigt sich aktuell vor allem mit Fragestellungen zu praktischen Gründen, moralischen Pflichten und Supererogation.

Intergenerational Distribution

A Matter of Justice

1. Introduction

Globalization has led to a rising interconnectedness between people and institutions all around the world: One may look at a simple ballpoint pen in one's fingers and see not only one's own but a multitude of hands in as many places designing its shape, producing the material it is made of, putting it together, building the machines that put it together, ship it, market it, sell it, and then, after some time, dispose of it.

However, the impact of our actions has not only increased through space but also time. As the issue of climate change clearly illustrates, that what we do now will influence the lives of our descendants to a great extent as well. That said, what is arguably even more important than our individual decisions are the economic, social, and political systems we foster as a whole, or in the words of John Rawls, "the basic structure of society" we adopt (Rawls [1971] 1999, 6). Hence, we might want to ask, how would the Rawlsian idea of just institutions look on a transgenerational scale? What are the demands of an intergenerational distributive justice?

In this paper, I understand a generation to mean the people alive together at one specific point in time. This differs from the common meaning of the term, yet it is in line with Rawls' use. In the context of his theory, this definition is useful since it circumvents having actual and possible people together in the original position. This, Rawls believed, would only obscure our intuitions. Later, Rawls states clearly that the original position features only contemporaries. He calls this "the present time of entry interpretation of the original position" (Rawls [1971] 1999, 120–121).

Alas then, at least according to Rawls' theory of justice, future generations propose a considerable problem. The lack of reciprocity between two generations, understood in the aforementioned manner, leads to the impossibility of cooperation or even simple interaction. Hence, it would appear the conditions of justice cannot possibly be satisfied. This is a prominent line of argument pursued in order to dismiss any intergenerational claims of justice based on a contractarian approach, for example, by David Heyd in his article *A Value or an Obligation? Rawls on Justice to Future Generations* (Heyd 2009, 168–169; see Barry 1989, 189; see Page 2007, 231–232). Even more discouraging might be the fact that Rawls himself only dedicated a few short passages in *A Theory of Justice* to the subject, namely chapters 44 and 45 (Rawls [1971] 1999, 251–262); passages that are then also very controversially discussed and led to a great amount of criticism such as the one Heyd has put forward.

Still, I on my part am of the opinion, that the impossibility of transgenerational distributive Justice is not a necessary conclusion from Rawls' theory of justice. Therefore, in this paper I want to argue against such an interpretation of Rawls, especially in respect to his idea of the basic structure of society. There are in fact different theories on how to understand the basic structure in Rawls' theory which in turn all lead to distinct conclusions about the scope of justice. The inexistence of cooperation does not equal a dismissal of any claims of justice for all of them. In fact, my aim is to demonstrate that intergenerational distribution *is* a matter of justice. For that purpose, however, let me first present the argument against that assertion in more detail.

2. A case against intergenerational justice

David Heyd argues in *A Value or an Obligation?* that the effects our institutions and political decisions have on our descendants do not in fact fall into the scope of justice laid out by Rawls in his *A Theory of Justice*. The argument mostly relies on the lack of reciprocity in intergenerational relationships. Whereas the current generation has substantial influence on the life prospects and well-being of the next, the latter due to the

unidirectional property of time has no possibility to do the same. It is only possible for contemporaries to engage with each other and while the past causes the future, the future can only, if at all, indirectly impact the past. Therefore, because there are no mutual relations, there can be no cooperation (Heyd 2009, 168). Or more specifically, Rawls' "circumstances of justice" do not hold in the intergenerational context, since they are in Rawls' own words "the normal conditions under which human cooperation is both possible and necessary" (Rawls [1971] 1999, 109).

Rawls divides such circumstances into two kinds: First, there are the objective circumstances such as a relative scarcity or physical and mental equality between people, concerning the environment in which cooperation is to take place (Rawls [1971] 1999, 109–110).

Second, there are the subjective circumstances characterizing the persons engaged in cooperation e.g. as having both conflicting as well as complementary interests. It is clear to see that not even the first such requirement Rawls mentions is satisfied in the case of intergenerational relationships, namely that "individuals coexist together at the same time on a definite geographical territory" (Rawls [1971] 1999, 110).

However, it is these background conditions that, by making cooperation feasible, give rise to the necessity of principles that govern the distribution of cooperatively produced goods, the benefits but also inequalities ensuing from working together. It is only under these circumstances that questions of justice may even arise. Thus, they can be said to define the role of distributive justice (Rawls [1971] 1999, 112).

As these circumstances do not hold in the intergenerational sphere, Heyd sees three possible ways to extend justice to future generations: first, to alter the idea of the circumstances of justice, second, to ground the idea of justice on non-contractarian principles, or finally, the position he favours, to admit that intergenerational relations are not subject to justice at all, but rather to morals or duties of another kind (Heyd 2009, 169).

He starts by criticising the several amendments Rawls made to the original position trying to accommodate for some kind of intergenerational saving principle, which would also meet the demands of distributive justice. In doing so, Heyd wishes to show that the circumstances of justice in fact cannot be altered in a satisfactory manner (Heyd 2009, 170–181).

My issue, however, is with his assertion that because the circumstances of justice are not fulfilled in the intergenerational relation, no claims of justice may prevail. It seems questionable that the scope of justice, or more precise, the boundaries in which the principles of justice hold should depend like this on the direct interaction of individuals. Questions of justice arise in societies; it is institutions, laws, distributions, and so forth that can be said to be just, not single people or acts. Such structures, however, are far more comprehensive than a group of people at a certain point in time, these structures at least are intergenerational.

For that reason, let us have a closer look at how Heyd argues that relationships between two generations are not subject to justice. The questions to ask here are: Do intergenerational relations really fail to satisfy the circumstances of justice by their lack of mutuality? And do the circumstances themselves as a matter of fact really define the scope of justice?

3. Cooperation: the sole ground of justice?

Although Heyd, in his article, proposes a few ideas on how to model a kind of reciprocity or cooperation between generations, he rejects them all as unsatisfactory (Heyd 2009, 174–175). I agree. However, I do not agree that therefore the scope of justice is limited only to the current generation. I think that, by considering cooperation as the single determinant in deciding whether or not the issue of justice arises, Heyd is simplifying matters considerably.

Let me refer here to the writing of Arash Abizadeh, *Cooperation, Pervasive Impact, and Coercion: On the Scope (not Site) of Distributive Justice.* Here, Abizadeh considers the basic structure argument of Rawls (Abizadeh 2007, 319). Since for Rawls "the primary subject of justice is the basic structure of society", it becomes crucial to define what exactly is meant by that basic structure in order to be able to assess the scope of justice (Rawls [1971] 1999, 6). As I wish to show next, it is by no means clear that the basic structure is limited to human cooperation. In fact, Abizadeh distinguishes three ways in which the basic structure of society might be understood: Firstly, as the institutions that govern social cooperation, secondly, as the institutions that have pervasive impact upon people's chances in life, and finally, as institutions that subject people to coercion (Abizadeh 2007, 319). Each of those interpretations in turn leads to a quite distinct understanding of the scope of justice; either as the set of people engaged in a scheme of social cooperation within the same institutions or pervasively impacted, respectively, coerced by the same institutions (Abizadeh 2007, 320). While Abizadeh's main objective is to argue for a cosmopolitan view on justice, a global scope of Rawls' principle, his arguments are also very interesting in the light of justice between generations.

I think it is quite evident that while it might be difficult or even impossible to argue for intergenerational cooperation, it doesn't seem too far-fetched at all to think of future generations as pervasively impacted or even coerced by the same institutions we are. Hence, in the following paragraph I will argue against the cooperation theory before I then want to propose why intergenerational relationships are a matter of justice in the other two theories.

It is not surprising that Heyd considers only the cooperation theory. Social cooperation is indeed fundamental for Rawls' approach to justice. For example, he characterizes the basic structure further as the "way in which the major social institutions distribute the advantages from social cooperation" (Rawls [1971] 1999, 6). Or as Heyd has rightly pointed out and as I have already addressed above, Rawls' descriptions of the circumstances of justice also feature the idea of human cooperation (Rawls [1971] 1999, 109; see Heyd 2009, 167–168). Additionally, with its contractarian approach and its use of a hypothetical social contract undoubtedly a lot of emphasis is placed on human interaction.

Even though Abizadeh manages to rather convincingly show that the scope of justice according to the cooperation theory cannot be only the people participating in cooperation, but rather more broadly any kind of social interaction, this still poses a problem to the intergenerational dimension (Abizadeh 2007, 331–333). I firmly believe that any theory trying to argue for even a weak form of reciprocity through time is not convincing. The future just cannot interact with the past. An action done now cannot be altered, reversed, or influenced by an action being performed at a later point in time. Of course, considerations of the future might impact our decisions in the present, but there is no direct causal way in which it might have an effect on it. Whereas, of course, the future depends on the past. Hence, I will try to refute the cooperation theory on other grounds.

While social interaction, at least according to the cooperation theory, might indeed be a necessary condition of justice, in the sense that only where there is social interaction, questions of justice arise, it does not necessarily follow that the scope of justice is similarly defined. If Rawls says the basic structure comprises "the way in which the main political and social institutions of society (a) fit together into one system of social cooperation, and the way they (b) assign basic rights and duties and (c) regulate the division of advantages that arises from social cooperation over time", then he first and foremost makes claims about its defining properties and not about its temporal duration (Rawls 2001, 10).

For example, it might be said that for a major social institution like the economy to exist, social cooperation is required. However, the same structure or organization could surpass several generations, each only allowing the interaction between its current members. Thus, if this were considered as one basic structure through time rather than several, and if the subject of justice is indeed the basic structure, it would then seem that the scope of justice might also extend to future generations. Rawls' assertion that it is the basic structure and not the individual actions which is the primary subject of justice highlights this point. He also states that an institution "exists at a certain time and place when the actions specified by it are regularly carried out in accordance with a public understanding that the system of rules defining the institution is to be followed" (Rawls [1971] 1999, 48). From that, it is quite difficult to argue that these specific actions should not be able to be carried out by several generations all in correspondence with the same system of rules.

For example, as long as public votes are carried out regularly and comply with the same constitutional requirements, a democracy lasts on. It is more likely to survive the demise of its original members than a change in the law.

Especially, if one takes into account points (b) and (c) in Rawls' description of the basic structure, it seems improbable that this would only include the people engaged in social interaction at the moment (Rawls 2001, 10). For both, the assignment of basic rights and duties and the division of advantages that arise from social cooperation usually encompass several generations. Further, the examples he mentions as major institutions: the political constitution, the principal economic and social arrangements, for instance, usually last far longer than one generation and even rely to some extent on a certain continuation, e.g., the concept of money, essential to our capitalistic economy, can only function if every user expects the currency to keep its value indefinitely (Rawls [1971] 1999, 6–7).

Hence, it can be noted that even when one adheres to the cooperation theory it is not that clear that the scope of justice should be limited to only one generation. The question is whether the structures the people cooperate in or the specific people cooperating is the crucial factor. Still, it is true that, from a strict cooperation theory standpoint, it remains the easiest to argue against intergenerational justice. Much more so than with the pervasive impact or the coercion theory. Yet, Rawls' justification why the basic structure should be

the primary subject of justice does not mention cooperation, but rather the profound effects it has on people's lives from start to finish (Rawls [1971] 1999, 7). Thus, this leads us away from the cooperation theory towards the pervasive impact conception of the basic structure.

4. Two alternatives

According to Abizadeh, two more ways to understand Rawls' basic structure argument remain: the pervasive impact and the coercion theory (Abizadeh 2007, 320). In the following two subchapters I will present these two approaches and argue why they would allow or even call for intergenerational justice. Therefore, let me begin with the broader of the two: the pervasive impact theory.

4.1 The pervasive impact theory

If one focuses mainly on Rawls' justification for the basic structure argument concerned with the profound effect our social institutions have, then, as Abizadeh has pointed out, the principles of justice should apply to all institutions that have a pervasive impact on a persons' life chances (see Abizadeh 2007, 342). He draws his reasoning mostly from Gerald A. Cohen, who has also directed attention to that very aspect of Rawls' description of the basic structure (see Cohen 1997, 20). While I, in this text, will not focus like Cohen has done on what that means in terms of global justice, I merely wish to show that most, if not all our major societal institutions, do have a pervasive impact on future generations. From the legal system and the constitution to how we use our resources, and the economic system we engage in; all those decisions substantially influence the life our descendants are going to live. As Rawls put it:

> "[T]he institutions of society favour certain starting places over others. These are especially deep inequalities. Not only are they pervasive, but they affect men's initial chances in life; yet they cannot possibly be justified by an appeal to the notions of merit or desert." (Rawls [1971] 1999, 7)

In my opinion the generation one is born in is morally quite arbitrary and can neither be justified by an appeal to merit nor desert, in addition to what kind of family or what socioeconomic circumstances one is born in within one generation. Nevertheless, it has a huge impact on the life one is going to live, depending to a large extent on the basic structure one finds oneself in. An economy based on exploitation of natural resources, on the destruction of the environment for example, quite clearly favours early starting places in the chain of generations over the later ones. The combination of moral capriciousness on the one hand and a large pervasive impact on the other leads to the need of some kind of justification for the way the basic structure distributes the life chances and possible benefits of cooperation. This could be achieved by showing that the basic structure is just for everybody who is impacted by it. Therefore, the scope of justice would become intergenerational.

As a conclusion, the principles of justice should apply to all institutions that have pervasive impact on a person's life chances regardless of whether the people involved cooperate or not (Abizadeh 2007, 342). Cohen remarks that Rawls empathically characterizes the basic structure as having such profound effects and that this is for Rawls the very reason it should be the primary subject of justice (Cohen 1997, 20–21).

This would mean a much broader scope of justice than under the assumption of a cooperation theory but also than under the coercion theory I will present next. Since coercion always has a pervasive impact on the people concerned, some of the following arguments could also be applied to the pervasive impact theory.

4.2 Coercion theory

According to the coercion theory the scope of justice is not defined by cooperation or the existence of a pervasive impact but the fact of state coercion (Abizadeh 2007, 345).

Michael Blake argues that the Rawlsian theory of justice should be interpreted as a way to assess whether coercive institutions are justifiable (Blake 2001, 265). As the basic structure is defined as the "arrangement of major social institutions into one scheme of cooperation", and Rawls further understands institutions as a "public system of rules" that guide positions and the rights and duties they entail, it seems probable that such a basic structure may always entail some kind of coercion (Rawls [1971] 1999, 47); at least in the way coercion is understood by Blake who sees it as an intentional action designed to replace the chosen option with the choice of another. It therefore expresses a relationship in which one party can dominate the other and violate their autonomy (Blake 2001, 272).

Accordingly, Blake's account of justice is based mostly on the idea of autonomy or rather how an act of violation against it might be justified. A system of rules can always be said to limit the autonomy of its subject

and hence being in need of justification. Blake thus interprets Rawls' theory of justice as giving an account of the "circumstances under which a coercive legal system could be justified to all those who live within it" (Blake 2001, 279).

This is quite similar to the pervasive impact theory, only that here it is coercive actions that need justification. Hence the principles decided on in the original position reflect the ways in which reasonable people are willing to give up their full natural autonomy for the possibility of social cooperation, rights, duties, the distribution of goods, and so on. The coercion must at least hypothetically find acceptance, even by those least favoured by it. This can be ensured by putting just institutions into place that satisfy the equality and difference principle developed by Rawls (see Rawls [1971] 1999, 53).

In *Political Liberalism* Rawls makes it quite clear that free and equal citizens also have an equal share in political and coercive power, since, in the original position, they would never agree to place matters like the constitutional essentials or the basic question of justice in the hands of a particular person or group of people. Therefore, such an authority could not be grounded in public reason (Rawls 1993, 61–62).

Blake takes this account to mean that Rawls intended his principles of justice only to hold within a set of individuals that share coercive political institutions, because only those need that kind of justification through public reasoning (Blake 2001, 287). Thus, he concludes that "coercion, not cooperation, is the sine qua non of distributive justice" (Blake 2001, 289).

However, if this interpretation of Rawls' theory is accepted, I think, it follows that the coercive nature of our institutions must not only be justified before our contemporaries, but also before our descendants within the same society. Whenever, we decide on a new legislation, a new rule, a new infringement on our autonomy we do that not only for ourselves but for all the future people being born into the very same legal system. For they will be coerced by its implications just as much as we are. They cannot simply appeal to the fact that it wasn't them who had set up the regulation, and even if they could, the form of valid appeal would be again limited by a constitutional structure built before them.

As Thomas Nagel, another proponent of the coercion theory, points out: Justice is in Rawls' understanding a specifically political value. It judges social institutions and not individuals and becomes only applicable where people have a distinct relation to each other, an institutional one. Hence, justice is something we only owe to those with whom we stand in a strong political relation. He describes it as an *associative obligation* (Nagel 2005, 120–121).

Even though, I am not sure whether to agree with Nagel on the implications such a conclusion has on the scope of global justice, in the case of intergenerational relations it clearly encompasses future generations. For he argues that what is unjust about arbitrary inequalities is not their existence per se but rather if they appear between fellow participants in a collective enterprise of coercively imposed legal and political institutions (Nagel 2005, 128).

And how is the date of birth not less arbitrary than race, sex, or the social position one is born in? Further, it certainly holds for future generations as well, perhaps even more so than for our fellow citizens, that without being given a choice, they are "assigned a role in the collective life of a particular society" (Nagel 2005, 129).

Additionally, as I have already argued above, they are just as much held responsible for obeying its laws and conforming to its norms. From that arises a request for justification according to Nagel. Otherwise, he claims it would be nothing more than "pure coercion" (Nagel 2005, 129).

Therefore, it would seem, in regard to the coercion theory as well, claims of justice arise between generations. If the scope of justice is understood to comprise the set of persons subject to coercion by the same institutions, then future citizens must be included as well (Abizadeh 2007, 320). This leads me to a short summarization of my argument so far and an outlook on its possible impact.

5. On widening the scope of justice

So far, I wish to have shown that if one follows Rawls' basic structure argument—the primary subject of justice is not the individual, but the way the major social institutions fit together and how they are arranged (Rawls [1971] 1999, 47)—then it must not follow that the scope of justice only covers the people engaged in some kind of social cooperation or interaction. As Abizadeh has pointed out, there exist as many as three distinct understandings of the basic structure in common literature which can be characterized as the cooperation, the pervasive impact and the coercion theory all with their own implications on the scope of justice (Abizadeh 2007, 319).

While I would agree with Heyd that the cooperation theory is the most unfavourable of the three in terms of arguing for intergenerational justice, it seems to me by no means the most convincing interpretation of Rawls' basic structure argument. Especially, not if cooperation is used as the single and exhaustive criterion to define

the scope of justice. As I have argued, a lot of our institutions, although cooperative in nature, surpass more than one generation and can undergo quite a lot of change concerning their members. Further, Rawls grounded his argument that the basic structure should be the primary subject of justice not on its cooperative features but rather on the impact it has on people (Rawls [1971] 1999, 7).

However, if one rejects the cooperation theory in favour of either a pervasive impact or a coercion-based reading of Rawls, then it is not so simple to dismiss intergenerational claims on justice anymore.

Hence, I think I have presented at least one argument to suggest that Heyd and others might have been a little bit too quick to dismiss an intergenerational scope of justice. I for my part think that the claims of justice *do* extend over the current generation. From here, I want to reflect on what exactly the implications of such a conclusion might be.

As the scope of justice defines the boundaries in which the principles of justice derived from the original position hold their claim, the next question is how these principles influence our relationship to the following generations. In other words, we must find a way to properly take into account the intergenerational dimension in questions of justice.

Which brings us back to Heyd's text, where he criticises, not unconvincingly albeit a bit prematurely, the ways Rawls tried to include the interests of future generations in his original position (see Heyd 2009, 172–176). Rawls' original position has undergone quite a few alterations through time: While, at first, Rawls presented the general assembly approach where people of different generations come together, he later rejected this theory in favour of the present time of entry one with the additional motivational assumption that people generally even in the original position, care for their direct offspring (Rawls [1971] 1999, 254–255). Later Rawls changed the motivational assumption of the "present time of entry" version for one of strict compliance, meaning that every previous generation has followed the principle of savings the contractors agree upon in the original position (see Heyd 2009, 172).

Yet, I prefer to focus on a different model for a transgenerational original position, the one proposed by Daniel Attas, since I consider his model to be more convincing, even though it shares a lot of the features of Rawls' "present time of entry" proposal. Rather than the constraint of full compliance, Attas introduces "the formal constraint of universality" (Attas 2009, 203). What he means by that, is that the principles derived from the original position must hold for everyone alike. They cannot be differentiated between generations, and hence all generations must follow the same principles. From that, it follows that principles are ruled out if they should be self-defeating, self-contradictory, or reasonable, only if others conform to a different one. Therefore, the difference to Rawls' idea is that not the knowledge in the original position is constrained but rather the available principles on the menu the contractors are able to choose from (Attas 2009, 203–204).

So, with this model the interests of all generations can be taken into account equally. Thus, now it remains to discuss what kinds of just principles might arise from such an original position.

To start, there is no reason to believe that the first principle, that the people would choose the most extensive equal scheme of basic liberties, should change only because now more people from different times have to be considered in the original position (Rawls [1971] 1999, 53).

Similar to Attas, I believe that the difference principle also remains more or less unchanged, except that now the maximin rule applies not only to the synchronic but also to a diachronic dimension. What is meant by that, is that we try to improve the position of our least advantaged contemporaries (although we don't know which generation we belong to) through both redistributing the goods available at the moment and through saving for the following generation (Attas 2009, 207).

Although it might be difficult to assess what the optimal just saving rate would be, because for that extensive hypothesising about future development would have to be done, I do believe it is easier to say what it most certainly would not be. In cases like climate change, where the exhaustive negative effects on future generations are almost undisputed and the costs of reducing greenhouse gas emissions in comparison seem bearable, it appears to be evident that the agreement in the original position would not be to postpone any further action.

Additionally, apart from any concrete principles that might follow from such a transgenerational understanding of justice, it would already be a big step forward to try to equally take into consideration the interests of future generations in the realms of politics, economics and so forth. Here, also the idea of Blake that the original position can be understood as a device to justify state coercion (Blake 2001, 284) might prove fruitful. One could then ask oneself: Would future generations consent to us applying that policy? Could we justify it in front of them? If we had to answer these questions with 'no' then the proposed policy should be rejected on the grounds of being unjust to future people.

6. Conclusion

The aim of this paper was to contest the view that intergenerational relations do not fall into the scope of justice in a Rawlsian sense, especially, to contest the kind of argument put forward by David Heyd. I did that by arguing in favour of either the pervasive impact or the coercion view on the basic structure of society as opposed to the cooperation one. Arash Abizadeh has suggested these three ways to interpret the basic structure in Rawls' theory and while Heyd only considers the latter, I wish to have shown that this means simplifying things considerably. For if either of the other two views is accepted, it becomes very difficult to argue why future generations indeed are not part of the set of persons distributive justice applies to. While it is of course still possible to hold on to the cooperation perspective, I hope to have succeeded at least in opening a route of further criticism against any simple dismissal of the intergenerational scope of justice in the Rawlsian sense. In addition, by widening the scope accordingly and by thus applying Rawls' principles of justice also to future people, a new need for justification arises. Justice now demands that we take into account the interests of future generations just as we would our own in the original position. Thus, a new perspective in political debate, decisions, and social cooperation is needed: A transgenerational one.

References

- Abizadeh, Arash. 2007. "Cooperation, Pervasive Impact, and Coercion: On the Scope (not Site) of Distributive Justice." *Philosophy & Public Affairs* 35 (4): 318–358.
- Attas, Daniel. 2009. "A Transgenerational Difference Principle." In *Intergenerational Justice*, edited by Axel Gosseries and Lukas H. Meyer, 189–218. Oxford: Oxford University Press.
- Barry, Brian. 1989. *Theories of Justice: A Treatise on Social Justice, Volume 1*. Berkeley California: University of California Press.
- Blake, Michael. 2001. "Distributive Justice, State Coercion, and Autonomy." *Philosophy & Public Affairs* 30 (3): 257–296.
- Cohen, Gerald A. 1997. "Where the Action is: On the Site of Distributive Justice." *Philosophy & Public Affairs* 26 (1): 3–30.
- Heyd, David. 2009. "A Value or an Obligation? Rawls on Justice to Future Generations." In *Intergenerational Justice*, edited by Axel Gosseries and Lukas H. Meyer, 167–188. Oxford: Oxford University Press.
- Nagel, Thomas. 2005. "The Problem of Global Justice." *Philosophy & Public Affairs* 33 (2): 113–147.
- Page, Edward A. 2007. "Fairness on the Day after Tomorrow: Justice, Reciprocity and Global Climate Change." *Political Studies* 55 (1): 225–242.
- Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press.
- ———. (1971) 1999. *A Theory of Justice*. Rev. ed. Oxford: Oxford University Press.
- ———. 2001. *Justice as Fairness: A Restatement*, edited by Erin Kelly. Cambridge, MA: Harvard University Press.

Dela Wälti (22) befindet sich zurzeit im siebten und voraussichtlich letzten Semester ihres Bachelorstudiums in VWL und Philosophie. Sie kann sich für viele philosophische Themen erwärmen; von theoretischer über praktische zu analytischer und kontinentaler Philosophie. Besonders interessiert sie sich aber für den Schnittpunkt mit anderen Disziplinen wie zum Beispiel der Politik und Wirtschaft.

Genocide: Essentialism and Identity

Critique on Lemkin's Groupism

1. Introduction

The massacres against the Rohingya people in Myanmar in the years 2016 and 2017 were devastating: 24 000 people were killed, 18 000 were raped, 700 000 fled to neighbouring countries (Global Conflict Tracker 2021). International organisations have expressed concern about the human rights situation in Myanmar (Human Rights Watch 2019). The UN sent special rapporteurs, who described the situation as "ongoing genocide threat", even in 2019 (OHCHR 2019). The international community has responded judicially to these crimes. Two proceedings were opened—one at the International Criminal Court ICC and one at the International Court of Justice ICJ. Negotiations at the ICJ began in November 2019. Aung San Suu Kyi is Myanmar's state counsellor, former human rights activist and was imprisoned under the military regime. She defended the military actions against the Rohingya before the International Court of Justice. She vehemently resisted the accusations of any genocidal action. San Suu Kyi does not deny that war crimes occurred—but not on the scale of genocide (OHCHR 2019).

This tragic contemporary example shows that the notion of genocide is not treated as other international crime like crimes against humanity or war crimes. It is a crime of a distinct kind, often accompanied with the expression "the crime of crimes" and therefore being the exemplary case of what lawyers call *malum in se* (Schabas 2012, 35), the maximus case of evil. From a legal point of view, this is misleading—there is no hierarchy between genocide and other war crimes, like crimes against humanity. Nevertheless, the term "genocide" has a worse connotation—that is why political leaders want to distance themselves from it as much as possible, as San Suu Kyi exemplifies.

The legal term of genocide was created after the Shoah by the Jewish-Polish lawyer Raphael Lemkin and consists of the greek word *genos* (tribe, clan) and latin *cide* (as an analogy to homicide) (Lemkin 1944, xi.). The targeted extermination of a human *genos* was what Lemkin wanted to make prosecutable under international criminal law. He defends a narrow essentialist group conception—Lemkin gives importance to the existence of groups as such, and not only to the individuals within the group. Following Lemkin, there is something additionally and essentially important about groups that cannot be broken down to the value of its parts. As the group as such is the target of a genocidal act, so must the group as such enjoy legal protection (Moses 2012, 22).

The target of this paper is twofold: For one, the legal concept of genocide will be presented. For another, the essentialism behind it will be critically analysed. In the first chapter, Lemkin's views on groupism will be discussed. Following this, various critiques of Lemkin's view are presented. There are those who criticise the scope of Lemkin's group conception, but there are also critiques who go further and criticise the underlying essentialism. These points are presented in chapters 3 and 4. The focus of this paper, however, will be on Hannah Arendt's critique. She defends the importance of group identity but opposes Lemkin's essentialism.

2. Lemkin's notion of genocide

In this chapter, the concept of genocide will be discussed. The legal innovation of the Genocide Convention will be presented, followed by the discussion of its underlying groupism.

2.1 The importance of groupism for genocide

Lemkin is often referred to as the "father of the Genocide Convention". He realized through the analysis of the Armenian mass atrocities that there was no legal term that condemned the crimes committed with the intention to erase national, religious, or national groups as such. After escaping the Shoah, Lemkin spent most of his life advocating the importance of the creation and ratification of the international convention on the crime he named

"genocide" (Schabas 2012, 103–106). In his monography *Axis rule in Occupied Europe*, he advocates the creation of a new legal concept that includes the following crimes:

> "*De lege ferenda*, the definition of genocide in the Hague Regulations thus amended should consist in two essential parts: in the first place should be included every action infringing upon the life, liberty, health corporal integrity, economic existence, and the honour of the inhabitants when committed because they belong to a national, religious, or racial group, and in the second, every policy aiming at the destruction or the aggrandizement of one of such groups to the prejudice or detriment of another." (Lemkin 1944, 93)

Firstly, what surprises most non-lawyers, is that in Lemkin's view, genocide does not necessarily require an act of mass killings. Genocide is an attack on the identity of a group, and not towards the life of individuals. What needs to be noted with special attention is that for Lemkin genocide is not an attack towards any group. It is limited to the concept of *national, religious, or racial groups.*[1]

Secondly, it is striking that Lemkin puts much weight on the intention to destruct a group. Not only the criminal act itself, the *actus reus*, must be present for the conviction genocide. Also, the malicious mindset, the *mens rea*, must be in place. Genocide is not a mass murder—it involves the specific intent to systematically exterminate a national, religious, or racial group, or in other words, a human *genos*.

These two aspects, the systematic attack to a group identity, and the intention behind the act, define in Lemkin's view the exceptional crime of genocide. Lemkin's linguistic innovation "reshaped the moral landscape of the world" (Luban 2006, 309) and has undisputedly a strong rhetorical power. Genocide is often identified as the crime of crimes, the embodiment of all evil.[2]

2.2 The cultural heritage argument

Lemkin gives high importance to group identities. For him, it is the group identity that mainly determines the individual's identity. Therefore, it is primarily the group, and not the individual *within* the group, that requires specific legal protection. As discussed above, following Lemkin's train of thoughts, genocide is a crime that cannot be broken down to cumulative incidents of murder and rape, but rather requires special attention to the perpetrators intention to damage a group as a whole[3]. Lemkin's approach can be best understood when analysing his argument about the importance of cultural heritage (Luban 2006, 309–318).

The first premise in Lemkin's argument is that cultural contributions are produced by groups, not by individuals. First, this means that group alliteration determines the individual's cultural contributions. Therefore, these contributions must rather be attributed to the group, than to the individual within the group. The individual's contributions can be accredited to the group it belongs to, rather than solely breaking it down to individuals' abilities. Second, this means that there are parts of culture that cannot be reduced to individual actions but are shared among a group. Language or religion serve as examples of so-called irreducibly social goods.[4]

The second premise defines that the sum of cultural contributions of all groups forms the cultural world heritage. In Lemkin's view, these contributions are shared among humankind. The more diverse in respect of groups the world is, the more diverse and richer is its cultural heritage.

From this, Lemkin concludes the erasure of one group decreases the cultural heritage of humankind. If contributions can be attributed rather to the group than to the individual, and if these contributions are shared among humankind, then the quantity of groups determines the quantity of world heritage – the less groups exist, the less world heritage is created or conserved. Lemkin claims that our world culture would be impoverished if, for example, the Poles had no opportunity to give the world a Copernicus, or if the Greeks could not give the world a Socrates. Likewise, Lemkin argues that cultural richness was lost through the extinction of indigenous peoples and the resulting loss of specific Mayan dialects (Moses 2012, 29). In other words, when the richness of the world diversity suffers, so does the cultural heritage. Following from this, by erasing an entire group, genocide leads to the decrease of cultural heritage. That is what makes it a crime against humankind as such – group destruction is a 'universal and enduring problem' (Moses 2012, 23).

The backbone of this argument is Lemkin's groupism– the view that groups are essential to human nature and

1 Lemkin's thoughts were not transferred 1:1 into the Genocide Convention. For instance, ethnic groups are also protected in the Genocide Convention. More detailed in Schabas (2012), 103–105.

2 Schabas, who gave his book "Genocide in international Law" the subtitle "The crime of crimes", probably contributed his share to this confusion.

3 Later, in the Genocide Convention, this has been adapted to the intention to destruct a group 'as a whole or in part'. More detailed in: Luban (2006), 312-313.

4 For a more detailed discussion of irreducibly social goods, see Taylor (1997).

have an intrinsic value (Moses 2012, 22). Lemkin argues that groups are intrinsically valuable to humanity because they produce culture and ensure that the spiritual resources of humanity are not exhausted (Moses 2012, 23). This view defends groups as internally homogenous, externally bounded actors with a common purpose that exist in a pre-political context. Their value arises from pre-political processes and is supported by the loyalty of group members, who become part of a group by birth (Luban 2011, 632).

If humanity suffers the effects of a group erasure in Lemkins sense, it seems reasonable to set up a legal convent that frames the crime of genocide and protects groups as main agents of culture. Insofar, the legal term of genocide does not advocate individual rights like not to be victim of murder in a mass atrocity but claims protection for groups not to be destructed. The individual is only subsidiary protected as a group member (Moses 2012, 31).

There is some critique on this argument. On the one hand, there is criticism that Lemkin's groupism does not protect all groups to the same extent, but only national, racial, ethnic, and religious groups. On the other hand, the strong notion of group identity itself can be criticised, as it is not individual people but groups which are primarily protected. These two critiques are presented in the following.

3. Critique on the exclusion of political groups

This definition of genocide is directed towards what at the time were considered national minorities—precisely, national, ethnical, racial, and religious groups. Political or social groups are excluded from Lemkin's argument, and consequently from the Genocide Convention. It has been shown and criticized by different authors like Luban (2006) or Moses (2012) that some groups enjoy a lower level of protection than other groups. For example, the 1965 mass killings against communists in Indonesia cannot be legally labelled as genocide because the attack was towards a political, and not a national, ethnical, racial, or religious group (Luban 2006, 317). This imbalance was tried to be equilibrated through the legal codification of crimes against humanity. Social and political groups are protected by the legal concept of crimes against humanity[5], but not by the Genocide Convention. There is no legal hierarchy between the concepts of crimes against humanity and genocide. However, there is a rhetorical hierarchy, as genocide is considered the worse crime in the common sense. This leads to victim groups expressing disappointment when international courts find a crime against humanity but not genocide, as happened for the sentencing of the Argentine military junta (Schabas 2012, 122).

The reason why political and social groups cannot invoke the Genocide Convention, is because Lemkin classifies them as being of less value to humanity. As discussed in the second premise of the cultural heritage argument, for Lemkin, groups are valuable to humanity because of their possible or actual contributions to world heritage. He denies this possibility to political groups. As it is shown above, national, ethnical, racial, and religious groups exist in his view pre-politically. Someone's national, ethnical, or racial, and in some cases even the religious, group affiliation is determined by birth and cannot be changed. In Lemkin's view, only groups affiliation that is stable over a long period of time and that can hardly be changed deserve further protection, because only they contribute to individual's identity as well as to the cultural heritage of humankind. They are the roots on which political organization can elaborate and have therefor an essential value for the world community. Therefore, in Lemkin's view, national, ethnical, racial, and religious groups deserve protection, while political groups do not (Schabas 2012, 113). Nevertheless, this leads to a legal imbalance between different groups which are legally protected in different ways—either only by the term of crimes against humanity, or additionally by the term of genocide.

4. Critique on group rights

The criticism is thus that not all groups are protected by the Genocide Convention. But the critique on Lemkin's concept of genocide[6] goes further than only criticizing its scope and exclusiveness. From a liberal point of view, two assumptions of Lemkin's groupism can mainly be criticized: First (1), the legal supremacy of group rights over individual rights, and second (2), and in relation to this, the assumed primordial conception of group loyalty.

(1) Individual freedom is the foundation stone of liberal theory. Groups should only be protected as conglomerate of individuals, but without its own agen-

5 In 1998, the Rome Statute of the International Criminal Court ICC was adopted, and now it defines crimes against humanity as a "widespread or systematic attack directed against any civilian population, with knowledge of the attack" (ICC Statute Art.7). The victim group must not necessarily be homogeneous and constant over time, in contrast to the Genocide Convention. Nor is protection categorically denied to social or political groups. All crimes that are subsumed under genocide are also covered by the penalties for crimes against humanity. For more details about the codification and the legal scope, see: Richard Vernon, (2012), 231–49.

6 Here, I will focus on Lemkin's view on genocide and groupism and not on the legal concept codified in the Genocide Convention 1948 or the Rome Statute 1998.

cy. Liberal thinkers oppose to Lemkin's worldview in which internally homogenous groups, and not individuals, constitute the nucleus of examination (Luban 2006, 317–320).

(2) From a liberal point of view, the assumed group loyalty can be criticized as it diminishes individual freedom. For Lemkin, group membership is determined by birth. From that moment on, loyalty to the group is presumed. Due to the essential requirement of a pre-political loyalty to groups, individuals do not have the freedom to choose their own identity. They are bound to a group identity by birth and enjoy legal protection widely because they are part of a specific group (Schabas 2012, 115–116).

In summary, the predominance of group rights over individual rights is criticised. There are also concerns that individual freedom will be compromised because in Lemkin's world view, group identity is defined by birth.

One might therefore conclude that the notion of genocide should be abolished and replaced by a more neutral legal term such as crimes against humanity, which does not give priority to group identity. However, there is a third stream of thought that criticises both Lemkin's groupism and liberal individualism. Hannah Arendt, as a political philosopher, has contributed to the discourse. In the following, her position shall be presented.

5. Arendt's critique on group essentialism

Similar to Lemkin, Hannah Arendt's life was deeply impacted by the crimes of the Shoa—she had to flee to the U.S., because as a Jewish philosopher, she feared antisemitic prosecution in Nazi Germany. Her work is furthermore influenced by the political situation in Europe during the first half of the 20th century. In her monumental work *The Origins of Totalitarianism* ([1951] 1962), Arendt analyses crimes in totalitarian states within the context of an unprecedented societal moral collapse, in which law becomes crime and crime becomes law. Therefore, in totalitarianism, crime is a matter of duty (Arendt [1951] 1962, 460–480). She became world famous for her analysis of the Eichmann Trials in Jerusalem. In her work *Eichmann in Jerusalem* ([1963] 1965), Arendt analyses that Eichmann's malice is of a banal nature. Therefore, his crimes cannot be reduced to a subjective *mens rea.*[7] Following Arendt, in times of totalitarian persecution, moral and legal categories of political life are confounded. (Arendt [1963] 1965, 150).

Arendt devoted her work partly to the question of the significance of group identity in international contexts. This also becomes particularly evident Eichmann in Jerusalem (Arendt [1963] 1965). In her view, the subject matter of International Criminal Law is humankind as a whole. A crime against humanity offends all of humankind because it denies the root conception of humanism, mainly, that all human beings belong to a shared commonwealth (Luban 2011, 626). Her generic label for any attack against humankind is the crime against humanity, which is the unwillingness to share the earth with the people who inhabit it. Genocide, in her view, is a paradigmatic crime against humanity. Arendt defines genocide as an attack to human diversity as such—not as defined in law, but in a literal sense insofar that genocide offends all humanity and therefore, anyone who commits genocide is a hostes humani generis, an enemy of all humankind (Luban, 2015, 308).

This can, at a first glance, remind us of Lemkin's cultural heritage argument.As I have shown above, Lemkin derives an intrinsic group value from the cultural heritage argument. In his view, groups have an intrinsic value because they determine largely the individual's contributions—groups exist pre-politically and require the individual's loyalty. In Lemkin's view, since groups are the main producers of cultural heritage, they are valuable for all humankind. But in fact, Arendt's opinion is diametrically opposed to Lemkin's groupism. She disagrees that groups have an intrinsic value and that only therefore they deserve special protection (Luban 2015, 309). For her, the nucleus that deserves love, protection and care is not a group, but the individual. Nevertheless, she argues that group identity becomes important in times of persecution. This is reflected in her political theory of group identity that will be presented in the following (Luban 2011, 630–635).

5.1 Arendt's political theory of group identity

Arendt rejects the idea of an essentialist group identity. Nevertheless, Arendt states that describing identity in terms of group affiliation is on some occasions politically necessary. This is what Luban calls "political theory of ethnic identity" (Luban 2011, 633).[8]

The concept of group identity that Arendt advocates is twofold: On the one hand, she propagates that group identity is created in times of persecution. On the other, that self-attribution to a group identity can be an act of solidarity in the face of persecution.

7 For a more detailed reconstruction of her argument on the banality of the evil, see: Luban (2011), 635–641.

8 In my opinion, the term "political theory of group identity" would be more suitable. Not only ethnic identity, but any kind of group identity can be self-attributed out of political necessity to defend one's integrity.

5.1.1 Creation of group identity through persecution

Arendt's political theory of group identity is based on what Luban later will call "subjective identification" and rejects Lemkin's pseudo-objective, essentialist group identification. According to Arendt, group identity does not exist in itself, but emerges in times of political contestation —there is not something essential about group identity; there is no ontological essence. Group identity is created through social interactions: identity is a process of ascription and identification. While Lemkin wants to protect group identity because he sees an essential core in national, racial, or religious groups, Arendt argues that these attributes—racial, religious, national—are ascribed to groups. This becomes clear in her records of the Eichmann trial (Luban 2011, 635–641).

Eichmann was the first person prosecuted under the Genocide Convention. In his case, it was not difficult to identify the victim group: the Jewish people. Nevertheless, it was not the Jewish People who identified themselves as Jewish. It was the Nazi regime who set up rules that defined who would enter in the category of Jewish. Many persecuted people did not identify themselves as Jewish, did not practice Jewish religion, and assimilated themselves to Christianity. Nevertheless, the persecution turned them into Jews. The perpetrator's stigmatization of the group created the group identity through persecution. The victim may only perceive its artificial group identity through the world view of the perpetrator, which does not refer to a primordial group identity (Schabas 2012, 73–99). Following Arendt, in times of persecution, it is not the *effective* national, ethnical, racial, or religious group identity, which is under attack, but the *ascribed identity* (Luban 2011, 634–635).

5.1.2 Group identity and solidarity

Arendt, however, does not deny the importance of group identity as such. On the contrary, she argues that self-attribution can create intra-group solidarity. In her view, one can only defend oneself on behalf of the group under attack. In her essay *Humanity in Dark Times*, Arendt (1968) describes that for many years, she considered the only adequate reply to the question who she was, was to be a Jew. With this answer, she was acknowledging a political fact that considered the reality of persecution (Arendt 1968, 17–18). Here, the affirmative reference to a group identity is a speech of alliance or confrontation. Referring to the group identity which is under attack is not a purely descriptive statement. Describing oneself as Jewish in a totalitarian or post-totalitarian regime is an act of resistance. In Arendt's view, with the self-claiming of a group identity which is under attack, one cultivates solidarity *within* the group as well as compels others *outside* the group to acknowledge the persecutions and to situate oneself in relation to it. In contrast to render homage to pre-political group affiliation by showing loyalty, Arendt's act of identifying herself as part of the attacked group is a highly political act of solidarity. On the one side, Lemkin describes loyalty to a pre-political group identity. In contrast, Arendt describes solidarity with and within an attacked political group (Arendt 1968, 17–23, Luban 2011, 634–635).

5.1.3 The making of political groups

Having in mind the critique of the exclusion of political groups presented in chapter 3, Arendt disagrees with Lemkin's essentialist view that only certain groups can make reference to and are protected by the Genocide Convention. However, in contrast to the liberal tradition, Arendt does not deny any legal importance of group identity. For Arendt, groups have value through the political life they create—the importance of groups resides in the group's involvement in political activity. Therefore, Arendt strongly objects the exclusion of political groups from the Genocide Convention. As Luban argues, in a strong Arendtian sense, only groups that claim political space for themselves have a value as such and therefore deserve protection (Luban 2011, 635).

5.2 Summary of Arendt's argument

Lemkin ascribes a high intrinsic value to groups. Arendt contradicts this essentialist assumption but does not agree with the liberal criticism of group rights as such. She bases the first premise of her argument on the fact that group membership in a genocidal situation is not based on objective characteristics but is determined by the perspective of the perpetrator. Group identity is an act of ascription—there is no essence of being Jewish for example. The second premise is based on the fact that group identities ascribed to oneself are important—one can only defend oneself under the name of this attacked identity. Through this self-ascription to a group identity, one can build in-group solidarity. Arendt concludes from this that groups need special protection. However, Arendt does not grant this protection to all national, ethnic, racial, and religious groups, but only to those who step into the space of action as political actors (Luban 2011, 635).

6. Concluding remarks

Raphael Lemkin, who suffered greatly from the horrors of the Shoah through his biography, has dedicated his life to ensuring that these indescribable crimes are included in international criminal law. He created the concept of

genocide, which punishes those crimes that are motivated by the intention to destroy a national, ethnic, racial, or religious group. Lemkin ascribes a higher value to certain groups than the sum of their members. He bases this belief on the cultural heritage argument, according to which all contributions from all groups contribute to the cultural heritage of humanity and the loss of such a group would damage all humanity.

Lemkin did not assign a high value to political groups because he deems that their contributions are not valuable for the whole of humanity. This exclusion also entered the Genocide Convention and has ever since been criticized harshly. More deep-rooted criticism targets his essentialist, groupist world view, in which group affiliation is determined by birth and can hardly be changed. It can be highlighted that Lemkin fails to recognize that it is precisely this pre-political, essentialist group identity that fractionalizes humanity and thus can lead to segregation and, in the worst case, genocide.

Arendt is a strong critic of Lemkin's groupism. Firstly, she criticizes that Lemkin has not considered that it is not the objective group identities on the basis of which people are persecuted, but the one attributed to them. In her view of the world, national, ethnic, racial, or religious homogeneous groups do not exist *a priori*, but are only perceived during instances of persecution. Secondly, one can therefore only defend against persecution to the extent that one invokes the attacked group identity. Insofar, Arendt defends the legal concept of genocide. Groups deserve special protection, but she does not agree with Lemkin's justification for this, nor with the scope of the protected groups. For her, only groups that claim the political space between its members deserve legal protection through the concept of genocide—in other words, solely political groups.

This paper has attempted to present the problem of group identity in international law. However, it became apparent that a deeper level was touched upon in the process: Namely, the question of whether there necessarily needs to be an essence to group identity in order to protect it. This connects to contemporary debates from both feminist and anti-racist issues around the problem of non-essentialist anti-discrimination (Young 1994, Alcoff 2006, Haslanger 2012). In addition, the debate was opened on the distinction between solidarity and loyalty—where solidarity was thought as a non-hierarchical, non-essentialist form of group identity, while loyalty assumes essentialism. This is under-theorized at this point and is in need of further substantiation, for example through Jaeggi's (2001) or Khader's (2018) understanding of solidarity.

References

Literature

- Alcoff, Linda. 2006. *Visible Identities: Race, Gender, and the Self*. New York: Oxford University Press.
- Arendt, Hannah. (1951) 1962. *The Origins of Totalitarianism*. Cleveland: Meridian.
- ———. (1963) 1965. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: The Viking Press.
- ———. 1968. "On Humanity in Dark Times: Thoughts about Lessing." In *Men in Dark Time*. 3–31. New York: Harvest.
- Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press.
- Jaeggi, Rahel. 2001. "Solidarity and Indifference." In *Solidarity in Health and Social Care in Europe*, edited by ter Meulen, Ruud, Will Arts and Ruud Muffels. 287–398. London: Springer.
- Khader, Serene. 2018. *Decolonizing Universalism*. New York: Oxford University Press.
- Lemkin, Raphael. 1944. *Axis Rule in Occupied Europe*. Washington: Carnegie Endowment for International Peace.
- Luban, David. 2006. "Calling Genocide by Its Rightful Name — Lemkin's World, Darfur, and the UN Report." *Chicago Journal of International Law* 7 (1): 303–320.
- ———. 2011. "Hannah Arendt as a Theorist of International Criminal Law." *International Criminal Law Review Georgetown Public Law Research Paper* 11 (3): 621–641.
- ———. 2015. "Arendt on the Crime of Crimes." *Ratio Juris* 28 (3): 307–325.
- Moses, A. Dirk. 2012. "Raphael Lemkin, Culture, and the Concept of Genocide". In *The Oxford Handbook of Genocide Studies*, edited by Bloxham, Donald and A. Dirk Moses. 20–41. Oxford: Oxford University Press.
- Schabas, William. 2012. *Unimaginable Atrocities: Justice, Politics, and Rights at the War Crimes Tribunals*. Oxford: Oxford University Press.
- Taylor, Charles. 1971. "Irreducibly Social Goods". In *Philosophical Arguments*. 127–145. Cambridge, MA: Harvard University Press.
- Vernon, Richard. 2002. "What Is Crime against Humanity?" *Journal of Political Philosophy* 10 (3): 231–249.
- Young, Iris. 1994. "Thinking about Women as a Social Collective." *Signs* 19 (3): 713–738.

Materials

- 'OHCHR | UN Independent International Fact-Finding Mission on Myanmar Calls on UN Member States to Remain Vigilant in the Face of the Continued Threat of Genocide'. Accessed 20 January 2021. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25197&LangID=E.
- Global Conflict Tracker. 'Rohingya Crisis in Myanmar'. Accessed 20 January 2021. https://cfr.org/global-conflict-tracker/conflict/rohingya-crisis-myanmar.
- 'How a Peace Icon Ended up at a Genocide Trial'. BBC News. Accessed 20 January 2021. https://www.bbc.com/news/av/world-asia-50709830.
- Mapping Myanmar's Atrocities Against Rohingya. 'Mapping Myanmar's Atrocities Against Rohingya'. Accessed 20 January 2021. https://mapping-crimes-against-rohingya.amnesty.org/.
- Human Rights Watch. 'World Report 2020: Rights Trends in Myanmar', 16 December 2019. https://www.hrw.org/world-report/2020/country-chapters/myanmar-burma.
- International Criminal Court. Rome Statute of the International Criminal Court (1998). [ICC-Statute]

Sarah Heinzmann (26) studiert Philosophie im Master und verfasst momentan ihre Abschlussarbeit über die Kritik und Legitimität staatlicher Bestrafung. Sie interessiert sich für Kritische Theorie, (Queer-)Feminismus und normative Politische Theorie.

Affordances and the Normativity of Emotions

This contribution first appeared in: Synthese 194 (11), 4455–4476. (https://doi.org/10.1007/s11229-016-1144-7). Thanks to Springer for the permission to reprint this paper in meta(φ).

1. Introduction

Many current theories take emotions to be fundamentally embodied (e.g., Prinz 2004; Maiese 2011; Hutto 2012; Colombetti 2014). While these approaches differ substantially in the details, the term "embodied accounts" will be used as an umbrella term to cover all of these accounts insofar as they agree on the following claims: Embodied accounts of emotion (i) take bodily reactions to play a constitutive role in emotions, (ii) take emotions to be about core relational themes such as "being dangerous" or "being offensive," (iii) aim to replace vehicle-internalism (i.e., assumptions about inner representations realized and processed by the brain alone), or at least aim for minimalist vehicles such as perception-like states with non-conceptual content, (iv) aim to give a naturalist explanation of emotions, and (v) aim to externalize core relational themes (i.e., the intentional objects of emotions, such as "danger" in the case of fear). These general claims have become popular since they fit very well with an enormous amount of recent empirical evidence from various sources suggesting that there are patterned reactions of the nervous system connected with different types of emotions (Kreibig 2010), and that emotional expressions and bodily postures occur early in infancy (Reddy 2008) and across cultures (Tracy and Robbins 2007), and have homologies in animals (Clark 2010). Embodied accounts point out that bodily reactions are not simply an output or a random byproduct of an emotion but are rather constitutive of an emotion's intentionality. Put roughly, embodied accounts aim to understand emotions as evolved complex patterns of bodily reactions whose (biological) function is to respond to situations of urgent concern. The bodily reactions involved in emotions partly constitute the emotions' being meaningful, because emotions are the result of an adaptive history in which the bodily responses became reliable responses to urgent situations.

Contrary to behaviorism, embodied accounts point out that emotions are intentional and evaluative: they are about things that matter for us. Yet contrary to cognitivist theories, embodied accounts highlight that the emotions' "aboutness" should not be explained by assuming complex inner representations such as judgments with a conceptual content. Instead the bodily reactions themselves and the history of the interaction between a skillful body and a structured environment constitute the meaningfulness of emotions.

As I will argue in the following, embodied accounts have not been paying enough attention to the normativity of emotions in all its aspects but should do so in order to give an adequate characterization of the phenomenon. Yet embodied accounts are constrained in the way that they can account for the normativity of emotions by the naturalist, externalist, and anti-vehicle-internalist claims they embrace. In the following I begin by sketching the claims that embodied accounts rely on. I then explain what the normativity of emotions consists in and point out why embodied accounts have not paid sufficient attention to the explanation of the phenomenon. I then propose my own approach on which we should conceive of the intentional objects of emotions in terms of affordances, where these affordances are part of a value-rich environment that the skillful organism is situated in. I argue that such an approach is committed to a certain form of normative realism but that this normative realism fits well within

a naturalist framework. I also suggest that affordances can be described as standing in complex relations to each other and the organism, and that these relations can (at least in part) account for what has been labeled as the emotions' being subject to rational norms.

2. Embodied accounts

Embodied approaches respond to a steadily growing body of evidence from developmental psychology, ethnology, and behavioral studies in animals suggesting that the expressive patterns and bodily postures associated with certain emotions such as pride, jealousy, embarrassment, and shame might be present in apes, infants, and across cultures.[1] Humans show stereotypical bodily reactions and expressions from a very early age: we swell with pride and hang our heads in shame; our hearts race in fear and our blood boils in anger. In the most general sense, embodied accounts conclude from these studies that emotions can best be described as embodied evaluations that evolved as direct responses to situations of a certain urgency for the organism.[2] The general claims that unite embodied accounts can be described in the following way:

(i) *Embodiment* Bodily reactions play a constitutive role in emotions. The increase of heartbeat, the release of adrenaline, and the tensed muscles that are part of fear form a pattern, and this pattern is not just some random output of an essentially cognitive emotional reaction. On the contrary, the pattern of bodily reactions is the result of an adaptive history in which it gained the function of constituting the emotion's "aboutness."

(ii) *Core Relational Themes* Embodied accounts agree with traditional cognitivist accounts that emotions are intentional and evaluative. To account for the particular "aboutness" of emotions, many embodied accounts adopt Lazarus's (1991)view that each emotion has a certain core relational theme it is about: fear is about something's being dangerous, anger is about something's being offensive, and so on.

(iii) *Anti-Vehicle-Internalism* Embodied accounts generally aim to replace the traditional view of cognitive processes as consisting of complex inner representations realized by the brain alone with approaches that tend to see cognitive processing as being constituted by the interplay between the skillful body and the structured environment. Some accounts completely deny vehicle-internalism for large areas of the mental. They deny that internal representations play any role in the realization of perception, sensation, or emotion. Dynamical accounts even tend to deny that the vehicle/content distinction is a useful distinction at all (e.g., Gallagher 2008; Chemero 2009; Hutto and Myin 2013; Colombetti 2014). Other accounts prefer to define representations in a more modest way by speaking of minimal or action-oriented representations that are not realized by the brain alone (Clark 1997).[3]

(iv) *Naturalism* Embodied accounts widely agree on methodological and ontological naturalism, i.e., they aim to explain emotions in continuity with the results on emotions from the sciences and further develop a picture of intelligent organisms as the result of an evolutionary process. Methodological and ontological naturalism both find support in the wide range of evidence that

1 See e.g., Draghi-Lorenz et al. (2005), Reddy (2008), Tracy and Robbins (2007), Clark (2010).

2 It is of course still highly controversial to what degree emotions are constituted by bodily reactions that have their origin in evolution. That cultural differences play an important role in emotion is beyond doubt. Yet even scientists such as Lisa Feldman-Barrett, who doubt that emotions can be individuated with regard to the patterns of bodily arousal they involve, would not deny that bodily arousal plays a central role in emotional coping strategies and that some basic bodily ingredients that combine to constitute different emotions can be identified (see e.g., Barrett et al. 2015). So today, even those who admonish theorists not to overestimate the role of bodily arousal make a much more differentiated claim than traditional cognitivists, who treat bodily arousal as a meaningless causal output or contingent byproduct of an emotion (e.g., Lazarus 1991; Nussbaum 2001). Still the debate about the individuation of emotions is not to be settled on empirical grounds. For the purposes of this paper, I will follow those who think that emotions can be individuated with regard to bodily reactions without further argument, but I argue for an embodied view in detail elsewhere (Hufendiek 2016).

3 Prinz (2004) is an exception here, since, strictly speaking, he is a vehicle-internalist. In arguing against cognitivists that emotions are perception-like rather than judgment-like Prinz suggests that the vehicles of emotions can be understood as being simple and non-conceptual in format. But for Prinz emotions are perceptions of bodily arousal and the vehicles of these perceptions are situated in the brain alone. The following discussion of the problems that anti-vehicle-internalists face when it comes to the normative dimension of emotions do include Prinz's approach, though, since these problems are largely a result of denying that emotions are *complex* inner representations, such as judgments with a conceptual content.

speaks to panculturally occurring features of emotions and a continuity among emotional reactions in humans and other animals.

(v) *Externalism* Embodied accounts (by and large) are committed to diachronic environmental externalism (DEE) and synchronic environmental externalism (SEE). In a nutshell, DEE claims that organisms typically develop the ability to respond to certain types of objects only if there is an adaptive pressure to do so. This implies that the object has to be ontologically prior (in some sense to be specified) to the response that is set up to respond to it. SEE is the claim that no complex, world-representing inner machinery is needed in order to interact with the world in intelligent ways, *because the world is out there* and organisms are well-adapted to it. This implies that we have to think of the object of the response in question as being present and causally effective under normal conditions.

The commitment to externalism (v) is central to why I claim that embodied accounts of emotions need to be more explicit about their ontological commitments. I will therefore develop this claim further in what follows.

3. Externalism

For the externalism claim, as well as for the other claims, it is certainly true that when it comes to the details, authors differ on what it means to embrace externalism: within the teleosemantic framework, Prinz (2004) is committed to a strong version of content externalism, while Hutto (2012) and also Colombetti (2014) would deny that emotions are representations that are subject to semantic norms in the first place. But all embodied accounts of emotions are equally committed to DEE, that is, to the claim that the environment takes on an active structuring role in the evolution of an emotion's intentionality. Biologically inspired versions of naturalism, and particularly the teleosemantic approach to emotion that Prinz is committed to, claim that organisms develop representational powers (or—in enactivist terms—the ability to enact meaning) in direct interaction with the world and in response to certain adaptive pressures exerted by their environments. As Prinz puts it: "Emotions are set up to be set off by core relational themes" (Prinz 2004, p. 66). This view depends on the claim that there were circumstances in the external world that our ancestors repeatedly faced, thus making it beneficial for these organisms to develop a representational mechanism to deal with them.

Yet, if what emotions respond to are core relational themes, the most straightforward conclusion would be to assume that core relational themes such as danger, loss, and so on must be objective occurrences with the causal power to create an adaptive pressure. If we say that emotions are "set up to be set off" by core relational themes, this implies that core relational themes are ontologically prior to the emotional responses they trigger. By affecting the organism in various ways, it was the core relational themes that created the adaptive pressure that resulted in the development of the emotional responses in question. Given that core relational themes such as "being dangerous to the organism" are usually described as things that are of value to the organism, this seems to commit one to normative realism.

Embodied accounts further rely on what I call *synchronic environmental externalism* (SEE). The claim is that organisms are situated in a structured environment that they directly interact with. Once we focus on direct forms of intelligent interactions with the outside world, we can refute approaches on which (complex) internal representations are something that stand in between the acting organism and the world and are realized by the brain alone. While controversy exists among different authors in the field concerning whether we need to eliminate representations altogether or simply to modify the meaning of the term (see claim (iii) above regarding anti-vehicle-internalism), there is broad consensus about SEE, i.e., the claim that no complex, world-representing inner machinery is needed in order to interact with the world in intelligent ways, *because the world is out there* and organisms are well-adapted to it. This appears to be a shift in the explanatory burden from the brain to the body and the world. The focus on the structure of the environment, and the organism's bodily skills to interact with it, allows us to explain and model intelligent behavior without necessarily assuming the involvement of complex cognitive processing. DEE and SEE are claims that complement the anti-vehicle-internalist claim that emotions do not essentially involve inner representations realized by the brain alone. The ontological commitments behind DEE and SEE will be central in the following discussion of how embodied accounts can explain the normativity of emotions.

4. The normativity of emotions

In *The Passions of the Soul*, René Descartes gives a definition of what envy is: "Envy is a kind of sadness mingled with hatred which results from our seeing goods coming to those we think unworthy of it. Such a thought can be justified only in the case of goods of fortune… But sometimes fortune gives advantages to someone who is really unworthy of them. Then envy stirs us up only because having a natural love for justice we are vexed that it is upheld in the distribution of these goods" (AT XI, 466 PA 182). Obviously Descartes thinks of envy as an emotion that entails a normative dimension, since it can be appropriate or inappropriate. Yet how can we describe an emotion's being appropriate?

Looking at the example of envy, if I get a job that somebody else has worked very hard for simply because I happen to belong to an influential family, then there is apparently more than one sense in which we can speak about envy as an appropriate response: envy is a *semantically* appropriate response insofar as the core relational theme of envy is in fact present. Envy is also an *evaluation*, or directed at something that is of (dis)value to us. Many authors make the point by simply saying that emotions are about values, thereby introducing some kind of value-realism into the story (e.g., Deonna and Terroni 2012). Yet the parsimonious naturalist might wonder whether we have to assume that emotions are about values at all. The most straightforward way to address this question is to go through Lazarus's description of core relational themes and see if one can find adequate redescriptions of these core relational themes in non-normative terms.

Take "being dangerous" as an example: if I'm afraid of a snake, and appropriately so since the snake really is dangerous for me, can't the snake's being dangerous for me be explained in chemical terms? A snake is dangerous because of its ability to bite and release venom into the blood. And the snake's being venomous can be analyzed in chemical terms. But although some instances of the core relational theme "being dangerous" refer to events that can be described in physical or chemical terms, the biological level is needed to get an adequate description of the properties *as they come to be represented* in core relational themes: fear doesn't represent the chemical structure of the snake's venom; rather, it represents the snake's being dangerous for our bodily wellbeing. Being dangerous, indigestible, or offensive are properties that cannot be reasonably explained with reference to physics or chemistry alone. To say that something is dangerous for something else introduces a kind of normativity into the story that can only be captured adequately within a biological framework. It makes sense to speak of "dangerousness" as a property that exists in relation to the organism only if the survival of the organism is introduced as a basic value in the first place. If survival is taken to be a basic value, then it makes sense to say that something is dangerous, indigestible, or offensive and to say that all these things are bad for the organism. Negative emotions represent something as being bad for the organism; more specifically, fear represents something as being dangerous; disgust as indigestible; and so on.

Yet with regard to envy and many other emotions, it is not only impossible to describe their intentional objects in non-normative terms, but it also sounds implausible that their intentional objects should be values with regard to biological standards. For emotions such as envy as well as guilt, shame, jealousy, and pride, the social context in which they occur is a *sine qua non*: these emotions (or homologous forms of these emotions) could not exist in nonsocial species outside of a social context and, what is more important, we need to refer to social rules and norms to spell out what they are about. Contrary to emotions such as fear and disgust, which can be described as being about value-properties whose being of (dis)value can be spelled out with reference to biological norms, jealousy is about being left out by others, guilt is about having transgressed a social norm, and so on. Outside of a social context, the emotions in question could not possibly have the intentional objects they have.[4]

To complicate matters even further, envy not only is unintelligible outside of a social realm that sets up the norms that envy is responsive to. Envy also seems to be subject to rational norms, insofar as we would consider a person irrational who tells us that she envies herself, just as we would find it irrational if a person who appropriately envied me for a job offer I received turned out not to be happy or relieved if she got the job in the end. This rational dimension of emotions has been analyzed in detail by Helm (2001). Helm takes emotions to be felt evaluations that concern things of import, where "each emotion in general imposes rational commitments on one to display a broader pattern of emotions with the same focus in the relevant (actual and counterfactual) situations" (Helm 2001, p. 70). The focus of an emotion, according to Helm, is the background object's

4 This is not meant to exclude that emotions such as jealousy and guilt could be results of evolution with a biological function. I would like to remain neutral on the question of whether they are. What I claim is that guilt and jealousy have violations of social norms as their intentional objects. But I do not make any claims about how and why they became the objects of these emotions.

having import for the subject, which makes intelligible the evaluation implicit in the emotion. If I happen to own a prize Ming vase that I am very proud of, this is an object of import for me. My fears of having cocktail parties in the house or my anger about somebody throwing a baseball close to the vase are both made intelligible in light of the import that my prize Ming vase has for me (see Helm 2001, p. 69). According to Helm, we need to think of emotions in a holistic framework, where belonging to a broader pattern of emotions with a common focus defined by focal commitments is a necessary condition on the appropriateness of particular emotions. While one does not have to buy into the details of Helm's approach, a comprehensive theory of emotions should be able to account for the phenomenon he is describing. Emotions and their intentional objects cannot be described as isolated responses with a biological function that explains what they are about. Rather the holistic character of emotions and the way that the pattern structure of emotions grants their rationality have to be accounted for.

Faced with these various aspects of the normativity of emotions, embodied accounts find themselves constrained in how they can account for the normative dimension of emotions: their commitment to naturalism constrains what embodied accounts can take norms and values to be; their commitment to externalism constrains embodied accounts to assume that the intentional objects of emotions can be defined (at least to some degree) independently of the emotional response; and their commitment to anti-vehicle-internalism constrains what embodied accounts can take the cognitive processes involved in emotions to be. Yet authors that address the normativity of emotions typically simply assume either that emotions are "about values" (Deonna and Terroni 2012), that the intentional objects of emotions can only be defined simultaneously with the emotion (e.g., Helm 2001, p. 63), or that emotions are judgments or judgment-like states with a conceptual content that are interrelated in content- and rationality-sensitive ways (Kenny 1963; Lazarus 1991; Nussbaum 2001; Solomon 2003). In the following I will introduce the concept of affordances and suggest that they can be understood as value properties in the environment of an organism, where some affordances owe their being of value for certain organisms to biological standards and others to social standards. This approach accounts for emotional objects and their normativity within a naturalist framework by taking emotional objects to be of instrumental value to the organism. Affordances are furthermore introduced as properties that stand in complex relations to each other and to the organisms that perceive them. A description of this structure can (at least in part) account for what appears to be the holistic character of emotions without thereby buying into non-naturalist claims about rational norms.

5. About affordances

I have argued so far that emotions have a complex normative dimension and that embodied accounts are restrained when it comes to accounting for this normative dimension since they are committed to naturalism, externalism, and anti-vehicle-internalism. While naturalism constrains how we can account for norms and values, externalism forces us to think of the intentional objects of emotions as being definable independently of the emotion, and anti-vehicle-internalism forces us not to think of emotions in terms of complex internal representations realized by the brain alone, even if what we want to explain is an emotion's being concerned with the violation of a social norm. In the remainder of the paper, I will argue that we can account for the normativity of emotions without abandoning the narrow framework of embodied accounts just described if we take emotions to be about affordances and take affordances to be partly constituted by relational properties that are of instrumental value to an organism and the evolved responses to such properties.[5]

Affordances, according to Gibson (1986), are features of the environment that exist in relation to the organism and its abilities. These features are of value to the organism, they offer possibilities for action, and they highlight things that should be approached or avoided. For a particular organism, certain fruits look edible; for persons of a certain size and shape, certain objects look sit-upon-able; and so on. Gibson assumes that perception always involves proprioception[6] and thereby has intentional objects that are fundamentally observer-relative although the external information that is picked up through perception is assumed to be real. A tree might look climbable to a squirrel but not to me, while the floor looks walk-upon-able to me but not to a fish. Perceiving an affordance means perceiving the environment with what it offers to the animal in question given its needs and skills. Things in the environment can be perceived as

5 I develop the theory of action-oriented representations elsewhere (Hufendiek 2016) and argue that it makes sense to talk about representations being about affordances, although Gibson himself was a radical anti-vehicle-internalist who denied *any* notion of internal representation. Here I will focus on the intentional objects of emotions and leave the question of whether emotions are representations and subject to semantic norms open.

6 "Egoreception accompanies exteroception, like the other side of a coin" (Gibson 1986, p. 126).

being of value, insofar as they match with the organism's needs and skills. The notion of value can be understood here as an instrumental notion. Things are not in any intrinsic or context-independent sense good or bad; they are good or bad for particular organisms in particular environments. But their being good or bad for the organism does not depend on the organism representing them (or responding to them) as good or bad. An animal in an unfamiliar environment might entirely lack the ability to represent some berries as indigestible. But these berries nonetheless have the property of being indigestible simply in relation to the organism and its needs. This is why we need to assume observer-independent relational properties as constitutive parts of affordances. Yet affordances depend on our skills to engage with the world as well, which makes them response-dependent at the same time. Both response-independent relational properties and the evolved responses to them are partly constitutive of emotional affordances. I will develop my view of the ontological status of affordances below.

An affordance theory of emotional objects fits nicely with an embodied account as it offers a good *prima facie* understanding of the relation among bodily reactions, world-directedness, and motivating potential in emotions: the emotions' "aboutness" is constituted by the bodily reactions they involve, and these bodily reactions can be about an external core relational theme or an affordance, because they were set up by evolution (or a learning history) to do so. The bodily reactions, by preparing the body for action, also give an observer-relative shape to the intentional object of the emotion as it is grasped by the subject. In the case of fear, rather than only representing something's being dangerous, we represent it as a danger-to-be-avoided.

Consequently, I suggest that we understand the ontological constitution of an affordance such as a "danger-to-be-avoided" to be two-fold: First, it necessarily includes a relational property such as "being dangerous" that (temporally speaking) precedes the affordance. Yet "being dangerous" is not sufficient for the existence of an affordance. Affordances are also response-dependent insofar as they only exist in relation to our skills. Being dangerous is a relational property that was present in the environments of our ancestors already before they were able to respond emotionally to dangerous situations. The pattern of bodily reactions that evolved in response to this relational property is a further constitutive part of there being an affordance: something that is not only related to us as living organisms, but also to us as agents with certain abilities. Emotions can be understood as representations of affordances that are not only *intentional* (about relational properties) but also *intensional* (involving a particular way of skillfully responding to and thereby grasping relational properties). But, crucially, emotions might be intensional without thereby entailing complex representations with conceptual content; rather, they may be intensional merely because the skillful bodily reactions that prepare one for action constitute a certain mode of presentation.

Current theories of affordance-perception are divided into radical anti-vehicle-internalist (e.g., Chemero 2009; Rietfeld and Kieverstein 2014) and minimal-representationalist accounts (e.g., Clark 1997). While I tend to side with the minimal representationalists and argue elsewhere for emotions as embodied action-oriented representations (Hufendiek 2016), I will not argue for emotions being representations in this paper. On the contrary, the view of affordances that I develop here is in principle adaptable to more radical anti-vehicle internalist frameworks. Crucially, talking about affordances as being intensional does not automatically imply representations. Hutto and Myin (2013) for example allow that things can feel or seem a certain way on the level of basic cognition, which is a way of accounting for skillful responses as constituting a mode of presentation in anti-representationalist terms.

In the following I will develop the ontological status of affordances further, as well as the instrumental value they have in relation to the organism, and the way in which affordances are intensional and stand in relations to each other that can be described as being subject to rational norms.

5.1 The ontological status of emotional affordances

With regard to its precise ontological status, Gibson's original term 'affordance' has been interpreted in several ways. Some argue that affordances are relational properties that establish selection pressure to which the organism then learns to respond (Reed 1996). Others take a different stance and argue that affordances are dispositional properties of the environment complemented by dispositions of the animal: for example, something can be walk-upon-able only if the animal already has the ability to walk, and therefore affordances cannot be seen as establishing selection pressure. In other words, on this view, the abilities presupposed by an affordance must be already in place for the affordance to exist (Turvey 1992; see also Chemero 2009 for a more detailed discussion).

So the puzzle that affordance accounts face is this: Affordances are partially constituted by properties of the environment, since they can come to be perceived

by the organism; yet affordances essentially contain a response-dependent element. On the one hand, it is certainly misleading to describe affordances simply as relational properties that establish a selection pressure, since they must be the result of a selection process. But, on the other hand, in order to explain the selection pressure that brings about an affordance we need to assume a kind of property that is of instrumental value for the organism independently of an organism's being able to respond to that property at all. I will argue in the following that affordances must be described as being partly constituted by relational properties that in turn cannot be characterized as response-dependent properties of any sort, since those relational properties must be in place *before* the organism acquires the ability to respond to them. But once an organism has acquired the ability to respond to dangerous situations, i.e. once the relevant action-tendency is in place, we can also speak of a response-dependent property being in place at the same time. "Being dangerous" is a relational property that can exist in relation to an animal even if the animal does not have the ability to detect or respond to this property at all. A "danger-to-be-avoided" is a response-dependent property that owes its identity to the organism's ability to respond to it. It is the existence of the relational property that sets up the selection pressure for the animal to develop the ability to detect it. So, strictly speaking, relational properties and response-dependent properties are both constitutive elements of an affordance.

To develop this idea further let me characterize what relational properties and response-dependent properties are and then explain how relational properties constitute affordances. A property is a relational property if an object has this property not intrinsically but only with regard to another object. Relational properties are objective in the sense that their existence does not in general depend on our representing them or in any sense responding to them. Response-dependent properties, on the other hand, are at least partly constituted by our representations or responses. Typical candidates for response-dependent properties are "being red" or "being funny," and there are different ways to unpack in what sense these properties are response-dependent.[7] Some accounts make the claim that the properties in question are entirely response-dependent, while others claim that the properties arise from an interaction of subject and world. In an embodied framework we need to assume that the objects of our emotions are at least partly constituted by relational properties that are not response-dependent in order to account for DEE and SEE. Let me explain.

Assuming that emotions are about objectively existing relational properties allows for the claim that these properties can be instantiated and individuated regardless of whether the organism is able to perceive them. A token of a relational property can be instantiated without the organism actually perceiving it, and there can be certain tokens of a relational property frequently present in the environment of a species, without any member of the species being able to represent this particular kind of token. For example, a new predator might enter the environment of a species and would constitute a danger even if no member of the species has the ability to represent this danger. The dodo is an example of a bird that was endemic to the island of Mauritius where it evolved in isolation from predators and as a result happened to be not only flightless but also fearless. The bird became extinct briefly after sailors introduced rats and other animals on the island in the seventeenth century that plundered Dodo nests. Dodos had no naturally evolved abilities to detect these predators and defend themselves or fly. Nevertheless it makes sense to say that the predators were dangerous for the Dodos. An account that takes "being dangerous" to be a relational property that can be instantiated independently of a species' being able to respond to that property has the advantage that it can account for the adaptive pressure that the property might place on the species.

While it makes sense to say that *certain kinds* of dangerous situations only start to exist at some point in the history of a species, it is hard to imagine a living organism in an environment where danger is never instantiated. Organisms that are alive but mortal and vulnerable are, by definition, organisms that can find themselves in danger. It is furthermore hard to imagine any historical or possible environment that entails nothing that could threaten an organism's wellbeing. "Being in danger" seems to be a property that deserves the name of a "core relational theme" in a very fundamental sense. Being alive, mortal, and vulnerable implies the possibility of being in danger. Accordingly, we find homologous forms of fear in animals such as zebrafish (Kalueff et al. 2012) and fruitflies (Gibson et al. 2015) that enables them to avoid dangerous situations. The same is true for disgust (Kelly 2011). As soon as there are living organisms in need of nourishing themselves who are able to absorb certain things but not others, they need to be able to distinguish the things that can be absorbed from those that cannot. "Being indigestible" can be seen as the rela-

7 See e.g., Kauppinen (2014) for an overview of response-dependent properties in metaethics, and Prinz (2007) for a naturalist view that assumes that moral properties are response-dependent.

tional property constituting the affordance that disgust is about, an affordance that could be labeled "indigestible-thing-to-be-rejected."

Still one may wonder why introducing relational properties and making such a fuss about them if in the end a response-dependent element needs to be added. Is a danger-to-be-avoided anything other than a response-dependent property? In short, yes. I insist on the distinction and on relational properties and response-dependent elements being both constitutive parts of emotional objects for the following reasons: I take relational properties such as "being dangerous" to be such that (a) they can exist in relation to an organism even if the organism doesn't even have the ability to respond to them, (b) these properties create the adaptive pressure for the response to evolve in the first place, and (c) these properties are of instrumental value (or disvalue) for the organism.

Thus we can see why it is necessary to include non-response-dependent relational properties in order to explain affordances. The standard motivation for introducing response-dependent properties is to avoid normative realism. However, the claim I want to make is that affordances do not owe their value to the way we respond to them. The things we (adequately) respond with fear to are dangerous for us and would be dangerous even if we lacked the responses in question. And we needn't avoid a commitment to instrumental value of this sort. Consequently, while an affordance as a danger-to-be-avoided does essentially include a response-dependent element, it essentially includes a non-response-dependent element as well.

5.2 The values and norm-violations perceived through affordances

Gibson thinks of affordances as objects that are of value to the perceiving organism:

> The perceiving of an affordance is not a process of perceiving a value-free physical object to which meaning is somehow added in a way that no one has been able to agree upon; it is a process of perceiving a value-rich ecological object. Physics may be value-free, but ecology is not (Gibson 1986, p. 140).

A naturalist approach can account for these values, since they can be understood as instrumental values, defined with regard to biological norms or standards such as the survival of the organism. As has been suggested above with regard to emotions such as fear and disgust, one might be happy to say their objects are of disvalue to us with regard to biological standards.

Yet, as has been suggested above, with regard to many emotions it looks implausible to determine their intentional objects in terms of bodily needs or biological values in the first place. For emotions such as guilt, shame, jealousy, envy, and pride, the social context in which they occur is constitutive for their intentional objects. We need to refer to social rules and norms to spell out what they are about. Contrary to the objects of fear and disgust, properties whose being of (dis)value can be spelled out with reference to biological norms, the object of jealousy is "being left out by others," guilt is about having transgressed a social norm, and so on. Outside of a social context, the emotions in question could not possibly have the intentional objects they have.

Saying that the object of guilt or jealousy is fundamentally social might seem to suggest that these emotions do not have adaptive functions, and (given some kind of biosemantic or biosemiotic framework) this might suggest that the content or object of these emotions is not determined by a biological function but rather by the social function of the emotion. Yet this does not follow from the claim that these emotions only can have intentional objects in a social context. Consider homologous forms of shame in apes. These are most adequately described as responses to rule-violations concerning one's status inside a rank hierarchy. While detecting a rule violation committed by oneself with regard to one's status involves representing an intentional object of a complex social nature, the ability to respond to such objects could still have an adaptive function. With regard to most social emotions, there is not enough evidence (yet) to make definite claims about the origin of these emotions, so the question of whether social emotions have adaptive functions or social functions must be left open. Yet even if we accept that we cannot come up with clear decisions on whether these emotions are at bottom subject to biological or social norms, we have to explain how the intentional objects of these emotions are constituted on an ontological level, since it is one thing to claim that "being dangerous" is a relational property but another thing to claim that "having violated a social rule" is a relational property that exists independently of our responses to it.

A central claim of this paper is that both of these types of relational properties, and in particular relational properties concerning violations of social rules, can exist independently of whether we represent them or not. As we will see, the properties in question can exist even in a species that is not able to detect them. I

will use guilt as a first example. Guilt can be seen as an embodied reaction to a certain kind of norm violation. The norm itself is established in the social environment through a collective practice and the mutual acceptance of that norm in the practice.[8] There are many rules that we make up as we go along in our social interactions. Rules can get established as conventions that the members of a social group follow without being explicitly represented beforehand. It is not a necessary criterion for a rule to be in place that somebody represents it and then purposefully establishes it.[9] It is a criterion for a rule that it is followed and can be followed by members of a social group. But people can establish these rules without the intention of doing so and might or might not come to represent these rules explicitly later on. Rules are visible in social contexts in the form of re-occurring patterns of behavior and in people's sanctioning behavior when a rule is not followed. Re-occurring behavior patterns are grounded in the dispositions of people to do similar things under similar conditions.

In the case of guilt there are typical manifestations of guilt-relevant norm transgressions. A caregiver's stern face or raised voice might, for example, indicate to an infant not yet able to understand language that she has violated a social norm. If a raised voice or stern face are reliable signs of having transgressed a social rule or norm, it can come to be a locally recurrent source of natural information through which the infant can detect that she has violated a rule or norm.[10] The normative property would in that case be detected through the facial expression or the tone of voice. It would supervene on the situation the caregiver is complaining about and could be grasped through natural information such as a stern face that frequently co-occurs with that property in the environment of the infant in question. While having transgressed a norm is the relational property that guilt is about, the relevant action tendency is to-make-amends-for. The affordance that guilt is about is therefore a transgression-of-a-rule-to-make-amends-for. Of course, this is just a tentative suggestion for how to think about the relational properties that social emotions appear to be about. A more detailed account would require a developed social ontology of rules and rule violations and how we are set up to deal with them, in order to explain emotional content within a naturalist externalist framework.

But it should be clear that in principle such an account assumes that (a) the norm violations that these emotions are about can be described as social constructions that we collectively make up without being aware of it. This is an explanation of the norms in question that fits within a naturalist framework. Furthermore it allows us to say that (b) the relevant relational properties can be in place prior to our ability to respond to them or represent them, which does justice to the proposed externalism including DEE and SEE. Finally, (c) the emotional responses to these properties can be understood as action-oriented representations that directly respond to recurring natural signs such as the stern face of a caregiver and grasp the relational property in question by perceiving the facial expression. This allows us to say that guilt is about a norm violation without embracing the claim that this presupposes an explicit representation of the self in relation to others and an explicit representation of social norms. In other words, we do not need to embrace vehicle-internalism in order to explain guilt.

A further example should clarify these points. Consider studies that suggest that shame occurs in apes as a rank-related emotion that motivates them to show subordinate behavior towards higher-ranked animals (Clark 2010). Baboons establish complex social rank hierarchies in their interactions and can come to behave accordingly in their social environments. The social structures in question can be understood as networks of social relations between the animals. Being a parent of, being higher-ranked than, or being a permanent grooming partner of another are relational properties that animals can have *qua* occupying a place in the social order (and not *qua* collectively representing that somebody is a parent or a grooming partner).[11] Social relations are constituted through re-occurring patterns

8 See Rietfeld (2008) and Rietfeld and Kieverstein (2014) for an approach to affordances that are the result of our social practices or ways of living. See also H.L.A. Hart and his famous distinction between habits and rules. Hart proposes several conditions for a rule being present, including (i) regularity of behavior, (ii) a standard of criticism, (iii) a tendency to criticize for violations of the rule, and (iv) felt bindingness of the rule (Hart 1961).

9 Though this is the way that Searle would want to have it in his social ontology (2010). The reason not to follow Searle here is that an account such as Searle's, which makes collective representation of something *as* something the constitutive principle of the social world, cannot account for DEE and SEE when it comes to our representing social rules or the violation of them. To fit the constraints on an embodied account, we need to think of the social world, as it comes to be represented in emotions, as being ontologically prior to the emotional responses in question.

10 A notion of natural information that allows conventional signs to serve as natural information as well is developed by Millikan (2004)and Chemero (2009).

11 I follow Haslanger's (2015) description of social structure here. Such a description, again, has the advantage that it is not limited to Searle's view that social entities can only be constituted by our collectively representing them as such. Haslanger does not apply the notion of social structure to animals. But as far as I can see, there is no in-principle reason why this should not be possible.

of behavior. Practices relate the animals to each other and the material world. Given such a scenario, it makes sense to say that rank hierarchy as a social structure probably developed first and produced an adaptive pressure for members of the group to develop new skills to behave appropriately in the group. Shame in apes is thus an embodied action-oriented representation of the social affordance of a "status-rule-violation-to-be-hidden." Obviously the group of rule violations that can be adequately represented in shame by humans is at least a little broader, as shame in humans does not appear to be restricted to rule-violations that concern one's status. Yet the animal example nicely shows how in principle a complex arrangement of rules of behavior can be established without being collectively represented by the social group as such. The whole rank hierarchy of a group can be established through such behavioral reactions and does not need to be represented by any of the animals participating in the social system. Instead the animals make up the rules in the interaction and implicitly represent parts of the system through the others' behavior.

Such an account is not meant to cover the complex dimensions that shame-related reasoning can have in human adults. It is merely meant to cash out the content of shame understood as an action-oriented representation in a naturalist externalist context: Shame understood as an adaptive response that developed in a social context is about the violation of a social norm. Understood as an action-oriented representation, the object can be described as a rule-violation-to-be-hidden, which is a bit of a shortened way of putting it, since in shame it is the person who committed the rule-violation who feels the urge to hide herself and not the fact that she violated the rule.

A naturalist can integrate the intentional objects of these emotions into her ontology by taking them to be the results of a social construction. In particular, on the view I have developed here, their being real and causally effective depends on their being present in recurring forms of social behavior, and not on their being collectively represented as such.

So it seems that we have found an account that does justice to the fact that emotions are about values and that these values have to be response-independent, without making ontological assumptions that are not coherent with naturalism. Yet a non-naturalist might still be unsatisfied with the claim that all there is to the normativity of emotional objects is that they are appropriate with regard to biological or social norms. As Bennett Helm puts it, "[an] appeal to biological fitness presupposes rather than explains import [of emotional objects]. For food, water, and shelter, as instrumentally necessary for my (or my genes') survival, are worth pursuing only insofar as my life or my genes are worth preserving, and the worth of these has simply been presupposed rather than accounted for" (Helm 2001, p. 51). This leads Helm to conclude: "This all suggests that we should try to answer the modified Euthyphro question the other way around: things have import to us because we evaluate them as good or bad" (Helm 2001, p. 51).

In response to Helm, an embodied theorist can say the following. First, it is not an option for a naturalist externalist account committed to DEE and SEE to answer the Euthyphro question the other way around. The example of species being surrounded by dangerous objects they are not even able to detect or of the apes setting up social rules without needing to understand them suggests that Helm's approach might indeed be an inadequate way of describing the objects of our emotions. Rather, what is of import to us seems to be structured (at least to a significant degree) by relations between living organisms and their environment independently of their abilities to respond to these environments. Moreover, Helm's further conclusion that "pleasure and pain are, plausibly, such fundamental conations, automatically constituting their causes as good or bad" (Helm 2001, p. 52) simply replaces the grounding of good and bad in biological survival values and posits an ad hoc mechanism of mental value constitution. What Helm suggests here is not an option for a naturalist account if the values are supposed to be anything else then entirely subjective. So without purporting to resolve the long-running metaethical controversy between naturalist and non-naturalist approaches to value, I propose instead to judge the present approach by its explanatory power with respect to the objects in question, in particular by its ability to account for the normativity of emotional objects within a naturalist framework without ending up with inadequate descriptions of those objects and without adding elements to our ontology that cannot be accounted for in naturalist terms.

5.3 The intensionality of emotions

Emotions, I have argued so far, are about affordances such as a danger-to-be-avoided. The emotions' aboutness is constituted by the bodily reactions they involve. These bodily reactions are set up by evolution or a learning history to respond to relational properties such as "being dangerous." Assuming that emotions are about objectively existing relational properties allows for the claim that these properties can be instantiated and individuated

regardless of whether the organism is able to perceive them. As an externalist account, an embodied account is committed to the claim that an emotion is appropriate if the relevant relational property is actually present.

Yet emotions are not only about something, emotions also present their objects in a particular way. The way in which an emotional object is given to us is what is called the intensionality of emotions. According to the present approach, the pattern of bodily reactions that develops in response to a relational property turns this property into an affordance and constitutes its intensionality. An affordance is a property that is not only related to us as living organisms but also related to our abilities to respond to that property. We do not only see something as dangerous but as a danger-to-be-avoided. Fear is thereby not only *intentional* (about the relational property "being dangerous") but also *intensional* (presenting something's being dangerous as something that should be avoided). The embodied action tendencies involved in emotions can be described as "modes of bodily attunement" (Fuchs 2013) that determine the kind of access we have to the object in question and the way we feel motivated to act towards it.

Furthermore, these embodied modes of presentation explain why through fear we are not able to recognize dangerous objects in all their possible disguises. I am not afraid of the danger that the wolf in sheep's clothing constitutes, because I don't see him as a danger-to-be-avoided, but it would, in principle, be appropriate to be afraid. In the same way the dodo is not able to see certain predators as dangers-to-be-avoided, although it would be helpful if he could. Following this approach, emotions can be seen as intensional without thereby entailing conceptual representations that represent something "under a certain description." Rather, it is a bodily mode of urgent avoidance tendencies that constitutes the way an object is given to us in fear, and this also gives a hint as to why emotions are not entirely penetrable by rational thought. Fear does not always vanish when we judge something to be harmless, because the way something is presented to us in fear is via an embodied mode that is often better accessible through deep breathing or associations than through the cognitive judgment that everything is fine.

5.4 Interrelations among affordances

We can think of affordances not only as being externally given, but also as standing in relations to each other. Affordances constitute the structure of our environment and motivate one to behave in certain ways with regard to our needs and concerns. We are surrounded by things that are dangerous and by situations in which we could violate a norm. Moreover, affordances also stand in relations to each other and to us that determine which emotions are appropriate in which context. Many things can be dangerous and disgusting at the same time, and it is therefore common to feel fear and disgust at the same time. But the same objects do not merit fear and relief at the same time. Relief is rather an appropriate response to a situation where a danger has been removed. So it is not only the case that the intentional object of relief (as Helm would put it) is a backward-looking emotion that stands in a rational relation to a forward-looking emotion, namely, fear, but it is the presence of the intentional object of fear, the affordance of being a-danger-to-be-avoided, that makes fear appropriate when it is present, and that also makes relief appropriate if it is removed.

The bodily reactions that constitute relief are in an analogous way connected to those that constitute fear. The action tendency that comes with relief does not occur out of the blue. Rather, the feeling of relaxation that accompanies relief is a follow-up response to the tension that comes with fear, where one prepares for urgent action, whereas the feeling of relief brings the organism back to normal. The lesson to learn from this example is that the holistic structure that Helm analyzes might to a significant part be constituted by the relations in which affordances stand to each other and to us—again—independently of our being able to represent these relations. And what might look like rational relations between patterns of emotions and their objects might just as well be described as well-adapted skillful responses to situations our ancestors encountered over and over. So, if we ask why it is not rational to envy oneself or to feel fear and relief with regard to the same object at the same time, part of the reason must be that another's-good-to-be-obtained is not a property that one can possibly have oneself, because it is a property that can be had only in relation to oneself and thus can be had only by others. The objects of fear and relief are such that they cannot be instantiated in the same object at the same time. And this explains why it is never appropriate to envy oneself or to feel fear and relief about the same object at the same time.

6. Contemporary embodied accounts

To sum up, I will contrast my own embodied account with those I have mentioned so far. The present account is similar to Prinz's account in its taking core relational

themes to be relational properties. Yet while Prinz has the same externalist reasons to think of relational properties as being objective, he does not take them as constituting affordances and thus does not think of emotional responses as being intrinsically motivating for action (Prinz 2004). Prinz instead takes fear to be about something's being dangerous. Our being motivated to avoid that thing is constituted, on Prinz's view, by a further neural appraisal. But this is a surprising and seemingly ad hoc move on Prinz's part, in particular because there seems to be no empirical evidence grounding the assumption that there is any such neural evaluation taking place. Given that an embodied account has the advantage that it can describe the bodily arousal involved in an emotion as being constitutive not only for the emotion's intentionality but also for the action tendency that makes up its motivating potential. Furthermore, Prinz does not discuss whether the relational properties in question are of (dis)value to the organism. It is therefore unclear whether Prinz would be willing to accept a version of normative realism such as the one sketched above. Yet if he would not accept such a view, it would become unclear how he would account for the proposed externalism of core relational themes, given that core relational themes cannot adequately be described in non-normative vocabulary.

The account I have proposed here significantly differs from Hutto (2012) in its commitment to representations. While I am sympathetic to the claim that emotions are subject to semantic norms and that these norms can be understood within a biosemantic framework, I did not argue for that claim in this paper and the view on affordances I have developed here is in principle largely compatible with Hutto's view. One point of difference between our views, however, is Hutto's commitment to emotions being divided into basic and higher cognitive emotions, such that he takes them to belong to different psychological categories, where basic emotions are not representational while higher cognitive emotions are (Hutto 2012, p. 176). In Hutto's broader account, this distinction puts emotions like fear and anger into the basic part of the mind, while emotions such as guilt and shame are put into the contentful part of the mind. Such an approach has been defended by Griffiths (1997), but has been criticized by many authors since then (see e.g., Prinz 2004; Clark 2010; Colombetti 2014; Hufendiek 2016). As these authors show, the aim of understanding emotions as a single psychological category is backed up by plenty of current empirical evidence. It is a virtue of a naturalist approach to be able to account for the most recent empirical evidence with regard to emotions, and as far as I can see, the evidence largely speaks in favor of a unified picture of emotions.

Furthermore, regarding so-called higher cognitive emotions in infants and apes, it remains to be explained on a view like Hutto's how they can have emotions such as guilt and shame, without ascribing complex conceptual representations like an explicit understanding of social norms to them. On this score I submit that the present account does slightly better than Hutto's, insofar as I suggest that these emotions are representational, but then I explain why it suffices to presuppose embodied action-oriented representations in order to account for emotions such as guilt and shame. Hutto seems to embrace both naturalism and externalism, but he does not clearly stake out a position himself with regard to the normativity of the intentional objects of emotions. I have argued that emotional affordances are constituted by relational properties that are of instrumental value to the organism. Hutto is not very explicit on the ontological status of the properties in question, but expresses doubts as to whether it is necessary to assume relational properties in order to account for emotional objects (Hutto 2012, p. 178). The present approach argues that the ontological commitment to relational properties and instrumental norms does an important explanatory job for the externalist.

Many current approaches point out that emotions are situated in a particular environment and argue that the unfolding of emotions in a particular situation as well as their development is *scaffolded* to a large degree by the biological and social environment (Krueger 2014; Colombetti and Krueger 2015; Slaby 2014). These approaches all point in a similar direction as my own account and lay the foundation for my assumption that emotion-relevant affordances are present in our environment and are communicated to us through a wide variety of signs, ranging from facial expressions to written imperatives. At the same time, these accounts lack the ontological ingredients necessary to account for diachronic and synchronic environmental externalism. To be able to account for the normative structure of emotions, we should flesh out the claim that the world is its own best model by accepting the ontological commitments of the normative realism that is needed to account for core relational themes.

Some accounts do better in this respect than others. Enactivist accounts such as Colombetti's agree on the emotions' being valent for the organism. Yet valence is something that is enacted, i.e., only present due to the meaning-generating interaction of the organism with the world. As we have seen, this does not account for diachronic environmental externalism. So at the very

least enactivist accounts would have to be more explicit about the adaptive pressure that is prior to the valence of emotional responses (Colombetti 2014). Furthermore, the notion of valence is introduced by Thompson (2010) to highlight that organisms are adapted to respond to things that are of biological value for them. As we have seen, many emotions are about things that are of value for us with regard to social rather than biological norms. There are enactivist accounts that explain how sense-making is supposed to work in the social domain (e.g., De Jaegher and Di Paolo 2007). Yet so far these accounts have not been applied to emotions. So enactivist accounts would need to show that they are as well-equipped as the present approach to account for the objects of all emotions within a naturalist and externalist framework.

To do so, we need to make assumptions about the structure of the world we are surrounded by and our relation to it. We need to assume that things are of value in relation to the organism and that they thereby have the relational properties of being good or bad for the organism. We further need to assume that social groups can make up rules through re-occurring patterns of behavior that acquire an objective status. What good or bad means can be further specified according to the organism's needs and the social structure the organism is situated in. Some things are good because they are nutritious and others bad because they violate a social norm. Core relational themes are value properties that are of central importance for the organism's wellbeing and therefore give a basic relevance structure to the environment of the organism.

7. Conclusion

Embodied approaches to emotion have not paid sufficient attention to the normativity of the emotions and to the question of how to account for this normativity. This is a problem, because embodied accounts are constrained in terms of what kinds of explanations they can offer, given that embodied accounts embrace naturalism, externalism, and anti-vehicle-internalism.

To solve this problem, I have argued that we ought to take emotions to be about affordances. Affordances can be described such that they are constituted by relational properties that exist independently of an organism's being able to respond to these properties. By adopting this approach, such an account does justice to externalism (both DEE and SEE). Affordances are taken to be value properties, such that something's being a-danger-to-be-avoided implies that it is of disvalue for the organism with regard to biological standards, and something's being a rule-violation-to-make-amends-for implies that it is of disvalue for the organism with regard to either biological or social standards. In this way, my account offers an explanation of an emotion's being about value in the case of both lower and so-called higher order emotions. Finally, since the values in question are of an instrumental nature, they clearly fit within a naturalist framework and do not require additional ontological assumptions.

Given that even social emotions such as guilt and shame can be taken to be embodied, action-oriented representations that are about affordances, we can replace the vehicle-internalist assumption that these emotions entail inner representations with an account that describes the interaction of a skillful body and an environment to which the organism is adapted. Carefully describing the action tendencies involved in emotions gives us an embodied notion of the intensionality of emotions, that is, of how objects are given to us in emotions. This allows us to account for the fact that we do not respond with fear to every object around us that has the relational property of being dangerous. Finally, the fact that affordances are part of a complex environment and as such are related to each other and to us gives us an explanation of why certain emotional responses (such as envy) only make sense in response to others and why some emotional responses (such as envying oneself) never make sense, but are strictly irrational.

I therefore suggest describing emotions as being about affordances, where affordances are constituted by relational properties and owe their particular action-oriented form to our skillful abilities. Describing the environment as being filled and structured by such affordances means to describe it as an environment that entails value properties in relation to us and our needs and thereby motivates us to action. Emotional representations of such an environment can be appropriate or inappropriate; in particular, they can be about social rule violations taking place in this environment and can therefore also be subject to social norms. In order to fully account for the normative structure of emotions, we need not describe emotions as complex inner representations, but we do need to describe the interplay between a skillful body and a structured environment as an intelligent process.

Acknowledgements

I would like to thank Joshua Crabill, Markus Wild and two anonymous referees for their helpful comments on earlier versions of this paper. I would also like to acknowledge founding by the SNSF (Swiss National Science Foundation PP000P1_139037/2).

References

- Barrett, L. F., Wilson-Mendenhall, C., & Barsalou, L. (2015). The conceptual act-theory. A roadmap. In L. F. Barrett & J. A. Russell (Eds.), *The psychological construction of emotion* (pp. 83–101). New York: Guilford.
- Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47, 139–159.
- Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.
- Clark, A. (1997). *Being there. Putting brain, body and world together again*. Cambridge, MA: MIT Press.
- Clark, J. (2010). Relations of homology between higher cognitive emotions and basic emotions. *Biology and Philosophy*, 25, 75–94.
- Colombetti, G. (2007). Enactive appraisal. *Phenomenology and the Cognitive Sciences*, 6, 527–546.
- ———. (2010). Enaction, sense-making and emotion. In J. Steward, O. Gapenne, & E. Di Paolo (Eds.), *Enaction: Toward a new paradigm for cognitive science* (pp. 145–164). Cambridge, MA: MIT Press.
- Colombetti, G. (2014). *The feeling body. Affective science meets the enactive mind*. Cambridge, MA: MIT Press.
- Colombetti, G., & Krueger, J. (2015). Scaffoldings of the affective mind. *Philosophical Psychology*, 28 (8), 1157–1176.
- De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. An enactive account to social cognition. *Phenomenology and the Cognitive Sciences*, 6, 485–507.
- Deonna, J., & Terroni, F. (2012). *The emotions. A philosophical introduction*. New York: Routledge.
- Descartes, R. (1649/1988). *The passions of the soul*. In J. Cottingham, R. Stoothoff, D. Murdoch (Trans. & Eds.), *Selected philosophical writings of Renè Descartes*. Cambridge: Cambridge University Press.
- Draghi-Lorenz, R., Reddy, V., & Morris, P. (2005). Young infants can be perceived as shy, coy, bashful, embarrassed. *Infant and Child Development*, 14 (1), 63–83.
- Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: Form, content and function* (pp. 17–36). Oxford: Oxford University Press.
- Fodor, J. (2009). Where is my mind? *London Review of Books, 31*, 13–15. http://www.lrb.co.uk/v31/n03/jerry-fodor/where-is-my-mind.
- Fuchs, T. (2013). The phenomenology of affectivity. In K. Fulford, et al. (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 612–631). Oxford: Oxford University Press.
- Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies, 16* (3), 351–369.
- Gibson, J. (1986). *The ecological approach to visual perception*. New York: Psychology Press.
- Gibson, W. T., Gonzalez, C. R., Fernandez, C., Ramasamy, L., Tabachnik, T., Du, R. R., et al. (2015). Behavioral responses to a repetitive visual threat stimulus express a persistent state of defensive arousal in Drosophila. *Current Biology*, 25 (11), 1401–1415.
- Griffiths, P. (1997). *What emotions really are. The problem of psychological categories*. Chicago: Chicago University Press.
- Hart, H. L. A. (1961). *Theconceptoflaw*. London: Clarendon Press.
- Hart, S. (2015). *Jealousy in infants. Laboratory research on differential treatment*. New York: Springer. Haslanger, S. (2015). What is a (social) structural explanation? *Philosophical Studies*, 173 (1), 113–130.
- Helm, B. (2001). *Emotional reason. Deliberation, motivation and the nature of value*. Cambridge: Cambridge University Press.
- Hutto, D. (2011). Philosophy of mind's new lease on life: Autopoetic enactivism meets teleosemiotics. *Journal of Consciousness Studies*, 18 (5–6), 44–64.
- ———. (2012). Truly enactive emotion. *Emotion Review*, 4 (2), 176–181.
- Hutto, D., & Myin, E. (2013). *Radicalizing enactivism. Basic minds without content*. Cambridge, MA: MIT Press.
- Kalueff, A. V., Stewart, A. M., Kyzar, E. J., Cachat, J., Gebhardt, M., Landsman, S., et al. (2012). Time to recognize zebrafish 'affective' behavior. *Behaviour*, 149, 1019–1036.
- Kauppinen, A. (2014). Moral sentimentalism. Response-dependence supplement. In *Stanford encyclopedia of philosophy*. Accessed April 2016 from http://plato.stanford.edu/entries/moral-sentimentalism/supplement2.html.
- Kelly, D. (2011). *Yuck! The nature and moral significance of disgust*. Cambridge (MA): MIT Press. Kenny, A. (1963). *Action, emotion, and will*. London: Routledge.
- Kreibig, S. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84, 394–421.
- Krueger, J. (2014). Emotions and the social niche. In M. Salmela & C. von Scheve (Eds.), *Collective emotions* (pp. 156–171). Oxford: Oxford University Press.
- Lazarus, R. (1991). *Emotion and adaptation*. New York: Oxford University Press.
- Lewis, M. (2014). *The rise of consciousness and the development of emotional life*. New York: The Guilford Press.
- Maiese, M. (2011). *Embodiment, emotion and cognition*. Basingstoke: Palgrave Macmillan.
- Millikan, R. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.
- ———. (2004). *Varieties of meaning*. Cambridge, MA: MIT Press.
- Nussbaum, M. (2001). *Upheavals of thought. The intelligence of emotions*. Cambridge: Cambridge University Press.
- Parkinson, B., Fischer, A., & Manstead, A. (2005). *Emotions in social relations, cultural, group, and interpersonal processes*. New York: Psychology Press.
- Prinz, J. (2004). *Gut reactions. A perceptual theory of emotion*. Oxford: Oxford University Press.
- ———. (2007). *The emotional construction of morals*. Oxford: Oxford University Press.
- Reddy, V. (2008). *How infants know minds*. Cambridge, MA: Harvard University Press.
- Reed, E. (1996). *Encountering the world. Toward an ecological psychology*. New York: Oxford University Press.
- Rietfeld, E. (2008). Situated normativity: The normative aspect of embodied cognition in unreflected action. *Mind*, 117 (468), 973–1001.
- Rietfeld, E., & Kieverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, 26 (4), 325–352.
- Searle, J. (2010). *Making the social world. The structure of human civilization*. Oxford: Oxford University Press.
- Slaby, J. (2014). Emotions and the extended mind. In M. Salmela & C. von Scheve (Eds.), *Collective emotions* (pp. 32–46). Oxford: Oxford University Press.
- Solomon, R. (2003). *Not passion's slave. Emotions and choice*. Oxford: Oxford University Press. Tappolet, C. (2000). *Emotions et valeurs*. Paris: Presses Universitaires de France.
- Thompson, E. (2010). *Mind in life. Biology, phenomenology, and the sciences of mind*. Harvard: Harvard University Press.
- Tracy, J., & Robbins, R. (2007). The nature of pride. In J. Tracy, R. Robbins, & J. Tangney (Eds.), *The self-conscious emotions. Theory and research* (pp. 263–282). New York: The Guilford Press.
- Turvey, M. (1992). Affordances and prospective control: An outline of the ontology. *Ecological Psychology*, 4, 173–187.
- Watson, J. (1930). *Behaviorism*. New York: Norton.

Rebekka Hufendiek ist SNF-Eccellenza Professorin am Philosophie Institut der Universität Bern und leitet dort das Projekt „Explaining Human Nature. Empirical and Ideological Dimensions". Ihre Forschungsinteressen liegen im Bereich der Wissenschaftstheorie, der Philosophie des Geistes und der Anthropologie.

Was ist eine Frau?

Eine Kritik deskriptiver analytischer feministischer Definitionen des Begriffs *Frau* im Hinblick auf das Prinzip der Transinklusion

1. Einleitung

Dass trans-Frauen Frauen sind, lässt sich nicht nur mit moralischen und politischen Überlegungen, sondern auch mit metaphysischen Gründen rechtfertigen. Die Gültigkeit von Transidentitäten kann demzufolge durch die epistemische Erste-Person-Autorität legitimiert werden. Entsprechend lässt sich aus der Proposition, dass trans-Frauen Frauen sind, das Prinzip der Transinklusion formulieren. Eine deskriptive analytische feministische Definition des Begriffs Frau, die trans-Frauen nicht unter allen Umständen inkludiert, erfüllt das Prinzip der Transinklusion nicht und ist somit zu verwerfen. Ein feministisches Anliegen ist es, systematisch gegen die Unterdrückung aller Frauen, ob cis oder trans, zu kämpfen, ungerechte soziale Strukturen zu dekonstruieren, hinzuarbeiten zu einer gerechteren Gesellschaft, in der kein Individuum aufgrund seines Geschlechts marginalisiert, exkludiert oder diskriminiert wird. Entsprechend ist es eine Verpflichtung der feministischen Philosophie, eine deskriptive Begriffsklärung von „Frau" zu formulieren, die trans-Frauen nicht exkludiert. In der vorliegenden Arbeit werde ich die uniessentialistische Definition des Begriffs „Frau" von Charlotte Witt in Hinblick auf das Prinzip der Transinklusion untersuchen und kritische Einwände formulieren. Anschliessend soll eine eigene deskriptive Definition des Begriffs „Frau" dargelegt werden, die sich als zweiteilige Auffassung charakterisieren lässt. Diese Definition hält dem Prinzip der Transinklusion insofern stand, als dass die weibliche Geschlechtsidentität die hinreichende Bedingung erfüllt, um als Frau zu gelten.

Was heisst es eigentlich für *mich*, eine Frau zu sein? Eine gute Frage, die ich mir vor der intensiven philosophischen Auseinandersetzung so nie gestellt hatte. Die subjektive Einschätzung, was es für *mich* oder für dich heisst, eine Frau zu sein, ist eng mit der komplexeren metaphysischen Frage verbunden, was objektiv gesehen eine Frau ist. Berufen wir uns auf unsere Intuitionen, lässt sich feststellen, dass wir ganz unterschiedliche Auffassungen darüber haben, was Geschlecht ist. So sind einige Nicht-Philosoph*innen als auch radikale Feminist*innen der Meinung, dass es ausschliesslich biologische Sexualmerkmale sind, die Menschen zu Frauen oder Männern machen (z.B Sheila Jeffreys 2014). Andere vertreten die Position, dass das Geschlecht eine sozial konstruierte Kategorie ist, und entsprechend die Sozialisation ausschlaggebend ist, ob und wann ein Mensch der sozialen Kategorie ‚Frau' oder ‚Mann' angehört. Darüber hinaus zirkulieren in akademischen Kreisen ganz unterschiedliche Ansätze, wie Geschlecht definiert wird oder werden sollte. Entsprechend soll in der vorliegenden Arbeit die Frage, was eine Frau ist, auf die sokratische Art und Weise gestellt und aus der Perspektive der feministischen analytischen Philosophie diskutiert werden.

Die philosophische Auseinandersetzung in Bezug auf die Frage, was Geschlecht ist, ist ein Projekt der feministischen Metaphysik, und zwar aufgrund der Relevanz der Frage für die feministische Theorie einerseits sowie der Perspektive bzw. der Herangehensweise andererseits und nicht zuletzt der Methodologie, die feministische Interessen widerspiegelt (vgl. Haslanger 2007). Entsprechend lässt sich die Frage aus unterschiedlichen Perspektiven und mittels unterschiedlicher Methoden bearbeiten: Die metaphysische Frage, was eine Frau ist, lässt sich demzufolge in sprachphilosophische Fragen umformulieren: Was bedeutet das Wort „Frau"? Wer fällt unter den Begriff der Frau? Was ist der semantische Inhalt des Begriffs „Frau" (vgl. Saul 2012)? Während sich das Konzept der Frau mittels unterschiedlicher Methodologie und Perspektiven untersuchen lässt, referiere ich in meiner Auseinandersetzung auf

beide Fragen gleichermassen, wobei ich die Position vertrete, dass der semantische Inhalt des Begriffs „Frau" mit den metaphysischen Annahmen bezüglich der Frage, was eine Frau ist, einhergeht. Eine Semantik von Geschlechterbegriffen, die nicht auf einer Metaphysik des Geschlechts basiert, wäre eine inhaltslose Semantik. Entsprechend habe ich den Anspruch an eine aussagekräftige Definition des Begriffs „Frau", sodass sie mit den metaphysischen Annahmen, was eine Frau ist, übereinstimmt.

2. Terminologie

Die Terminologie, die in den Diskussionen über Geschlechts- und Trans-Themen verwendet wird, ist umstritten (vgl. Bettcher 2014). Gerade deshalb ist es unumgänglich, zu Beginn einige Schlüsselbegriffe zu erläutern. Diesbezüglich definiere ich den für diese Arbeit im Fokus stehende Begriff der Frau vorerst nicht ausführlich. Während im englischen Sprachgebrauch eine linguistische Unterscheidung zwischen *sex* und *gender* gemacht wird, gibt es im deutschen Sprachgebrauch keine synonyme Übersetzung der inhaltlichen Unterscheidung zwischen dem biologischen und dem sozialen Geschlecht. Aus Klarheitsgründen übernehme ich jedoch die englische Terminologie, indem ich das biologische Geschlecht mit „das Sex" und das soziale Geschlecht mit „das Geschlecht" bezeichne. „Sex" referiert auf bestimmte biologische Eigenschaften, die durch die Geschlechtschromosomen X und Y festgelegt werden. Darauf folgen die Ausbildung von Hoden und Eierstöcken (Gonaden) und schliesslich der übrigen Sexorgane wie Penis und Vulva (Genitalien). Ein Neugeborenes wird als männlich oder weiblich charakterisiert, wenn es entweder männliche oder weibliche Sexmerkmale besitzt. Die Dudendefinition des Begriffs „Frau" liegt nicht weit davon entfernt: Eine Frau ist laut dem Duden eine erwachsene Person weiblichen Geschlechts, wobei „*weiblich*" mit dem gebärenden Geschlecht angehörend definiert wird.[1]

Dass die Kategorien 'Frau' und 'Mann' sozial konstruiert sind, griff Simone de Beauvoir in ihrem 1949 erschienen Werk *Le Deuxième Sex* auf eine eindrückliche Weise auf: Laut ihrer Auffassung wird man nicht als Frau geboren, sondern *on le devient* (1949, 334). Die Schriftstellerin und existentialistische Philosophin ruft in ihrem zweibändigen Essay Frauen dazu auf, sich nicht mit ihrem deklarierten Schicksal abzufinden und das zweite (frz. le deuxième) Geschlecht nach dem männlichen zu sein. Darauffolgend lieferte die Metatheorie des Sozialkonstruktivismus, die auf dem 1966 erschienenen Buch *The social construction of reality* von Berger und Luckmann basiert, Grundlage dafür, die Geschlechtskategorien insofern aufzufassen, als dass sie durch soziales Handeln erzeugt, konstruiert, institutionalisiert und erhalten werden (1966). In diesem Zusammenhang gewann der englische Begriff „gender" an Bedeutung: „Gender" (lat. genus = Gattung) wurde ab 1975 unter anderem vom Sexualwissenschaftler John Money und der Feministin Gayle Rubin etabliert, von Judith Butler in der Queer-Theory weiterentwickelt und später ins Deutsche übernommen, um auch hier, wie zuvor schon im angloamerikanischen Kulturraum, eine sprachlich erweiterte Unterscheidung zwischen juristischem, sozialem und biologischem Geschlecht einzuführen. Allerdings wird der Begriff „Gender" in diesem Kontext im deutschen Sprachraum meist mit „sozialem Geschlecht" übersetzt und dient zur analytischen Kategorisierung.

In der vorliegenden Arbeit verwende ich die Adjektive „maskulin" und „feminin", um auf die sozialen (stereotypischen) Geschlechtseigenschaften, Geschlechterrollen und sozialen Positionen zu referieren. Dass Frauen Röcke und lange Haare tragen, wäre entsprechend nicht eine weibliche, sondern eine feminine Eigenschaft. Maskulinität und Femininität definiere ich als kausal sowie auch substantiell sozial konstruiert. Die Frage, was genau das biologische sowie soziale Geschlecht sind, und ob eine Unterscheidung überhaupt sinnvoll ist oder ob das biologische Geschlecht nicht genauso sozial überformt ist, sind aktuelle Kernthemen der deskriptiven analytischen feministischen Philosophie (vgl. Mikkola 2010), die jedoch an dieser Stelle nicht weiter ausgeführt werden.

Im Folgenden wird zudem nicht explizit zwischen *transsexuellen- und transgender*-Identitäten unterschieden, wobei sich transsexuelle Menschen eher mit einem anderen biologischen Geschlecht und Transgender sich mit einem anderen sozialen Geschlecht identifizieren. Vielmehr spreche ich allgemein über *trans-Menschen* als Menschen, die *trans* sind. Es ist von einem trans-*Mann* die Rede, wenn der betreffende Mensch mit weiblichen Sexmerkmalen geboren wurde, sich jedoch als Mann (FTM: female to male) identifiziert. Von einer trans-*Frau* spreche ich, wenn der betreffende Mensch mit männlichen Sexmerkmalen geboren wurde, sich jedoch als Frau (MTF: male to female) identifiziert. *Cis* verwende ich, um all diejenigen Menschen zu beschreiben, die weder trans noch nonbinär sind, d.h. all diejenigen, bei denen das bei der Geburt zugeschriebene Sex mit der Geschlechtsidentität übereinstimmt. Folglich

1 Wörterbuch Duden. Wörter „Frau" und „weiblich". https://www.duden.de, zuletzt abgerufen am 31. Juli 2021.

ist eine cis-Frau genau dann eine cis-Frau, wenn sie sich mit dem bei der Geburt zugeschriebenen weiblichen Sex identifiziert.

Nicht zuletzt soll die empirische Tatsache betont werden, dass die hegemoniale, westliche Gesellschaft von einem heteronormativen, binären Geschlechtermodell geprägt ist: Die zweigeschlechtliche Einteilung in die Kategorien ‚Mann' bzw. ‚Frau' sowie die als selbstverständlich angesehene heterosexuelle Entwicklung konstituieren die sexuellen und geschlechtlichen Normen und somit die als „normal" bezeichneten Verhaltensweisen.

3. Prinzip der Transinklusion

In den letzten Jahrzehnten haben sich im deskriptiven analytischen feministischen Diskurs Philosoph*innen damit beschäftigt, eine Analyse von Geschlechterbegriffen zu entwickeln, insbesondere des Begriffes der Frau. Dieses Vorhaben stellt bis heute eine grosse Herausforderung dar, da es den Anschein macht, dass es keine essentielle Eigenschaft gibt, die allen Frauen gemeinsam ist. Wird die Frage, was eine Frau ist, im Sinne einer sokratischen Frage diskutiert, müssten notwendige und hinreichende Bedingungen angegeben werden, um den Begriff der Frau zu definieren. Da jedoch der Begriff „Frau" sowohl moralisch als auch politisch und gesellschaftlich eine unterschiedliche Relevanz und Bedeutsamkeit entfalten kann, steht bei einer Definition einiges auf dem Spiel: Je nachdem, wie „Frau" definiert wird, fallen trans-Frauen entweder unter den Begriff der Frau, oder aber sie werden marginalisiert und exkludiert. Unter Marginalisierung verstehe ich nicht die vollständige Exklusion. Vielmehr ist eine Beschreibung oder Charakterisierung der Kategorie ‚Frau' gemeint, die gewisse trans-Frauen als Grenzfälle einstuft und Frauen mit dem weiblichen Sex als Paradigma-Fälle im Gegensatz zu den trans-Frauen situiert. Folglich wird bei deskriptiven Projekten sowie Begriffsamelioration zu Beginn oftmals für ein Prinzip der Transinklusion argumentiert, wie unten genauer ausgeführt wird. Eine deskriptive Definition, die dieses Prinzip nicht beachtet, ist eine nicht adäquate Definition.

Das Prinzip der Transinklusion lässt sich in Bezug auf unterschiedliche Herangehensweisen bei der Begriffsdefinition auf mindestens zwei Arten interpretieren: Das Prinzip kann als moralische und politische Forderung aufgefasst werden. Das Prinzip der Transinklusion besagt demzufolge, dass eine Definition des Begriffs „Frau" trans-Frauen nicht exkludieren darf, weil die Konsequenzen des Misgenderns moralisch verwerflich sind. Strebt man ein deskriptives Projekt an, um „Frau" zu definieren, lässt sich aus dem Aspekt, dass Misgendering moralisch verwerfliche Folgen für trans-Frauen impliziert, jedoch nicht folgern, dass trans-Frauen Frauen sind. Mit der normativen Beurteilung der Folgen von Misgendering wird lediglich gerechtfertigt, dass es moralische Gründe gibt, dass trans-Frauen nicht in die moralisch falsche Geschlechtskategorie eingeordnet werden sollten.

Hinsichtlich der folgenden Argumentation fasse ich jedoch das Prinzip der Transinklusion im Sinne der zweiten Interpretationsmöglichkeit auf: Es gibt metaphysische Gründe dafür, dass trans-Frauen Frauen sind. Daraus wiederum lässt sich das Prinzip der Transinklusion ableiten: trans-Frauen sind aus metaphysischen Gründen Frauen und fallen entsprechend unter den Begriff der Frau. Die Beachtung dieses Prinzips ist genau dann erfüllt, wenn die Definition von „Frau" trans-Frauen inkludiert. Werden trans-Frauen vom Begriff „Frau" exkludiert oder marginalisiert, handelt es sich um eine Begriffsdefinition, die das Prinzip der Transinklusion nicht beachtet und somit zu verwerfen ist.

4. Uniessentialistische Definition des Begriffs „Frau"

Witt argumentiert dafür, dass das Geschlecht[2] uniessentiell für jedes soziale Individuum ist (2011, preface xiii). Jedes soziale Individuum ist demgemäss vergeschlechtlicht.[3] Dabei lässt sich Witt von der Theorie des Uniessentialismus leiten. Im Zentrum der von Aristoteles begründeten Theorie steht die Frage nach der Einheit und Organisation von materiellen Einzelteilen in einem neuen Individuum (Aristoteles 1959). So konstituieren laut der Theorie des Uniessentialismus beispielsweise Holz, Beton und Stahl genau dann die neue Entität ‚Haus', wenn sie die funktionellen Anforderungen an ein Haus erfüllen, das heisst, sie bieten z.B. Schutz für den Menschen. Im Vordergrund steht also die Frage, was die Einheit des Individuums erklärt. Damit unterscheidet sich die aristotelisch-inspirierte Theorie des Uniessentialismus von der Theorie des Identitätsessentialismus – die sich auf Kripke zurückführen lässt –, indem bei Letzterer die Frage im Fokus steht, was ein Individuum zu diesem Individuum macht. Laut der aristotelischen Theorie erklärt eine funktionale Essenz

2 Mit „Geschlecht" referiert Witt in erster Linie auf die binären Kategorien ‚Mann' bzw. ‚Frau'. Die Diskussion, ob Witt in ihrer Analyse auch von mehr als zwei Geschlechter ausgehen kann, führe ich im Kapitel 4.

3 Ich übersetze den englischen Ausdruck „to be gendered" mit „vergeschlechtlicht sein". Ist das Geschlecht qua soziales Individuum uniessentiell, ist jedes soziale Individuum notwendigerweise vergeschlechtlicht.

die existierende Entität, die über die konstituierenden Einzelteile hinaus existiert. Die funktionale Essenz spielt demzufolge eine unifizierende Rolle.

Laut Witt funktioniert das Geschlecht auf eine ähnliche Art und Weise: Es stellt das Prinzip der normativen Einheit bereit und organisiert, vereint und determiniert folglich die Rollen sozialer Individuen. Es ist wichtig, zu betonen, dass Witt das Geschlecht an die soziale Individualität knüpft: Geschlecht ist eine soziale Position, die jedes soziale Individuum essentiell besetzt, wobei die Position an die Reproduktionsfunktion *Engendering* geknüpft ist, die sozial vermittelt und von der biologischen Reproduktionsfähigkeit zu unterscheiden ist. Frauen empfangen und gebären, während Männer erzeugen. Engendering ist demgemäss die sozial vermittelte Form der biologischen Reproduktionsfunktion (Witt 2011, 37). Die sozial vermittelte Reproduktionsfunktion ist laut Witt die soziale Realisierung der menschlichen Reproduktion. Folglich ist ein soziales Individuum laut Witt genau dann eine Frau, wenn das Individuum die soziale Position [Frau][4] besetzt. Das Individuum besetzt laut Witt genau dann die soziale Position [Frau], wenn die sozial vermittelte Reproduktionsfunktion, die mit [Frau] assoziiert wird, sozial anerkannt wird.[5]

In diesem Zusammenhang ist die ontologische Unterscheidung zentral, die Witt zwischen der *Person*, dem *sozialen Individuum* und dem *Menschen* zeichnet: Während viele Philosoph*innen den ontologischen Unterschied zwischen Personen und Menschen anerkennen, vertritt Witt die Haltung, dass es noch eine dritte ontologische Kategorie gibt: das soziale Individuum. *Menschliche Wesen* sind laut Witt biologische, menschliche Organismen. *Personen* sind Individuen, die die Perspektive der Ersten-Person einnehmen können und die Fähigkeit zur Selbstreflexion sowie zur Autonomie besitzen. *Soziale Individuen* sind Individuen, die synchron sowie diachron soziale Positionen besetzen, soziale Rollen einnehmen und aufgrund ihrer Besetzung sozialer Positionen Subjekte sozialer Normativität sind. Diese ontologischen Kategorien sind weder äquivalent noch haben sie die gleichen Persistenz- und Identitätsbedingungen. Soziale Individuen existieren in Relation zur sozialen Welt sowie zum Netzwerk sozialer Positionen.

Witt fasst den Begriff der Frau insofern auf, als dass sie darunter eine soziale Position versteht, die mit sozialen Normen und Rollen verknüpft ist, die sich auf die sozial vermittelte Reproduktionsfunktion richten:

> *"My definition of gender—being a woman and being a man—ties these social positions to engendering; to be a woman is to be recognized to have a particular function in engendering." (Witt 2011, 39)*

Entsprechend lässt sich laut Witt eine deskriptive Definition mit notwendigen und hinreichenden Bedingungen wie folgt formulieren:

> *X ist genau dann eine Frau, wenn X ein soziales Individuum ist und [Frau] besetzt. X besetzt genau dann [Frau], wenn die sozial vermittelte weibliche Reproduktionsfunktion ‚empfangen und gebären' sozial anerkannt wird.*

Die soziale Anerkennung der sozial vermittelten Reproduktionsfunktion beinhaltet die Annahme sowie Anerkennung, dass X einen Körper hat, der diese spezifische Reproduktionsfunktion erfüllen kann: Wird sozial anerkannt, dass X eine Frau ist, impliziert das die Anerkennung, dass X einen Körper hat, der in der Lage ist, Samen in der Eizelle zu befruchten und ein Kind auszutragen. Zudem sind mit der sozialen Anerkennung der sozial vermittelten Reproduktionsfunktion auch soziale Normen assoziiert, die kontext- und kulturabhängig sind und die über die Zeit hinweg variieren können. Die metaphysische Auffassung von Witt, was Geschlecht ist, stimmt folglich mit den Geschlechtsidentitäten von cis-Menschen überein. Individuen, die mit den biologisch weiblichen Sexmerkmalen geboren wurden und sich als Frauen identifizieren, werden wohl in den meisten Fällen auch sozial als Frauen anerkannt: Der Körper mit den weiblichen Eigenschaften ist ausschlaggebend dafür, ob anerkannt wird, dass das entsprechende Individuum die sozial vermittelte weibliche Reproduktionsfunktion erfüllen kann oder nicht.

5. Einwände

Während jedoch bei trans-Menschen die biologischen Sexmerkmale und der Körper nicht mit der Geschlechtsidentität übereinstimmen, ist fraglich, wie die metaphysische Theorie von Witt mit den Erfahrungen von trans-Menschen vereinbar ist. Folglich diskutiere ich in diesem Abschnitt zwei mögliche Auslegungen von Witts Definition von „Frau".

5.1 Wechseln trans-Menschen ihr Geschlecht?

Gemäss der ersten Auslegung wäre Witts Theorie zufolge eine trans-Frau so lange ein Mann, bis sie in der Gesellschaft als Frau sozial anerkannt wird, d.h. bis

4 Mit [Frau] bezeichne ich die soziale Position der Frau. Entsprechend ist ein Individuum genau dann eine Frau, wenn es [Frau] besetzt.

5 Ich referiere auf *Reproduction* mit „biologischer Reproduktionsfunktion" und auf *Engendering* mit „sozial vermittelter Reproduktionsfunktion".

ihre sozial vermittelte weibliche Reproduktionsfunktion sozial anerkannt wird. Nach dieser Argumentation sind die Erfahrungen, Gefühle und Wünsche von trans-Frauen wie *Ich identifiziere mich als Frau / Ich würde gerne die soziale Position [Frau] besetzen* in Bezug auf die faktische Besetzung der sozialen Position [Frau] nicht determinierend. Die trans-Frau X ist erst dann eine Frau, wenn sie die soziale Position [Frau] de facto besetzt, wobei die Besetzung dieser unifizierenden Position durch die soziale Anerkennung der sozial vermittelten weiblichen Reproduktionsfunktion determiniert wird. Wird jedoch X in der Gesellschaft als ein soziales Individuum anerkannt, das die soziale Position [Mann] besetzt, ist X dieser Auslegung nach ein Mann und keine Frau, auch wenn sich X als Frau identifiziert.

Laut dieser Auslegung machen körperliche Operationen von Sexmerkmalen, die Einnahme von Hormonen und die äussere Erscheinung die Differenz bei trans-Menschen: Erst diese körperlichen Veränderungen führen dazu, dass eine trans-Frau nun als ein Individuum sozial anerkannt wird, das die sozial vermittelte weibliche Reproduktionsfunktion erfüllt und somit die soziale Position [Frau] besetzt. Laut dieser Auslegung findet bei trans-Menschen, die körperliche Eingriffe vornehmen lassen, ein Geschlechtsübergang statt: Der Übergang von Mann zu Frau bedeutet, dass das alte soziale Individuum, das die soziale Position [Mann] besetzt hat, aufgehört hat und dafür ein neues anfängt zu existieren, weil es nun die soziale Position [Frau] besetzt. Dabei bleibt die Person die gleiche, und nur das soziale Individuum macht diese Veränderung durch.

Die Tatsache, dass es soziale Individuen gibt, die weder die soziale Position [Frau] noch [Mann] besetzen, d.h. die Witts Definition zufolge weder Frauen noch Männer sind, führt dazu, dass Witt entweder dafür argumentieren muss, dass es sich bei diesen Individuen um keine sozialen Individuen handelt oder aber, dass es mehr als die binären Geschlechter Mann bzw. Frau gibt. Die Implikation, dass trans-Menschen, welche in der Gesellschaft weder als Frauen noch als Männer sozial anerkannt werden, keine sozialen Individuen sind, ist meines Erachtens unplausibel. Es scheint mir plausibler, dass es mehr als nur die binären Geschlechter gibt. Witts Definition könnte also auch so ausgelegt werden, dass trans-Menschen (hier trans-Frauen) weder Frauen noch Männer sind, sondern dass es ein drittes Geschlecht gibt, welchem sie angehören.

5.2 Haben trans-Menschen ein drittes Geschlecht?

Die Tatsache, dass es beispielsweise in Südostasien offiziell anerkannte dritte Geschlechter gibt wie die Hijra[6], die weder als männlich noch als weiblich definiert werden, lässt Witt folgern, dass trans-Menschen in den USA und Europa auf die Art und Weise charakterisiert werden könnten, als dass sie ihr Geschlecht nicht wechseln, sondern dass sie einem dritten Geschlecht angehören (2011, 40–42). Die Beschreibung der Hijra als drittes Geschlecht stimmt mit der Terminologie von Witt überein: Die Hijra werden weder als Frauen noch als Männer beschrieben, als auch bezüglich der Unfähigkeit, die sozial vermittelte weibliche oder männliche Reproduktionsfunktion zu erfüllen. Demzufolge handelt es sich bei den Hijra um soziale Individuen, die die soziale Position [Hijra] besetzen und demzufolge als [Hijra] sozial anerkannt werden, d.h. als soziale Individuen, die weder die sozial vermittelte weibliche noch männliche Reproduktionsfunktion erfüllen. Die Hijra gehören aber in Übereinstimmung mit Witts Terminologie nur dann einem dritten Geschlecht an, wenn sie auf die gleiche Art und Weise wie Männer und Frauen definiert werden, nämlich bezüglich der sozialen Anerkennung der Besetzung einer sozialen Position.

Eine deskriptive Definition von „trans“ im Sinne eines dritten Geschlechts könnte demzufolge wie folgt formuliert werden:

X ist genau dann trans, wenn X ein soziales Individuum ist und [trans] besetzt. X besetzt genau dann [trans], wenn die sozial vermittelte Reproduktionsfunktion ‚trans' sozial anerkannt wird.[7]

5.3 Ist das Prinzip der Transinklusion erfüllt?

In diesem Zusammenhang erachte ich jedoch folgende zwei Punkte als kritisch: Erstens stellt sich mir die Frage, inwiefern dafür argumentiert werden kann, dass die Besetzung der sozialen Position [trans] gleichermassen sozial anerkannt wird wie die Besetzung der sozialen

6 Das Wort „hijra“ ist ein Hindi-Urdu-Wort, abgeleitet von der semitisch-arabischen Wurzel hjr im Sinne von: „mit etwas brechen, verlassen, im Stich lassen, verstossen, auswandern, fliehen“. Es ist in Südostasien die Bezeichnung von Eunuchen, Intersexuellen und trans-Menschen.

7 Hier verwende ich die soziale Position [trans] für Menschen, die sich als trans-Frauen oder trans-Männer identifizieren. Es liesse sich jedoch diesbezüglich dafür argumentieren, dass sich die soziale Position [trans] des Weiteren in zwei unterschiedliche Geschlechter aufspalten lässt, beispielsweise einerseits als [MTF] und andererseits als [FTM]. Demzufolge könnte anhand der sozialen Position [trans] jedoch kein drittes Geschlecht abgeleitet werden. Vielmehr gäbe es entsprechend mindestens vier Geschlechter: [MTF], [FTM], [Frau] und [Mann].

Positionen [Frau] oder [Mann]. Die soziale Realität beweist das Gegenteil: Menschen, die sich nicht in die binären Geschlechterkategorien einordnen lassen, werden oftmals weder sozial anerkannt noch mit Respekt behandelt. Folglich ist es nicht einleuchtend, dafür zu argumentieren, dass trans-Menschen sogar als drittes Geschlecht sozial anerkannt würden. Viele Menschen bestreiten überhaupt die Existenz non-binärer Geschlechtskategorien. Um jedoch ein Geschlecht zu sein, sei es Mann, Frau oder Trans, ist die soziale Anerkennung eine notwendige Bedingung.

Zweitens ist meiner Meinung nach nicht nachvollziehbar, inwiefern sich die soziale Anerkennung der Position [trans] charakterisieren lässt, da sie bei den binären Geschlechtern mit der sozial vermittelten Reproduktionsfunktion eng zusammenhängt. Was genau beinhaltet die soziale Anerkennung der sozial vermittelten *trans* Reproduktionsfunktion? Inwiefern lassen sich trans-Menschen durch den Verweis auf eine Funktion bezüglich der menschlichen Reproduktion definieren? Wird die soziale Anerkennung insofern definiert, als dass ein soziales Individuum genau dann die soziale Position [trans] besetzt, wenn sozial anerkannt wird, dass dieses Individuum weder in der Lage ist, die sozial vermittelte weibliche noch männliche Reproduktionsfunktion zu erfüllen? Während beispielsweise die Hjira in Südostasien auf diese Art und Weise als dritte Geschlechtskategorie charakterisiert wird, werden die Erfahrungen von trans-Menschen nicht genug eingefangen: trans-Menschen sind sehr wohl in der Lage, eine sozial vermittelte Reproduktionsfunktion zu erfüllen. Trans-Menschen können Kinder erzeugen, gebären, fürsorglich aufziehen und erziehen etc. Denn nur der Gesichtspunkt, dass ein Individuum trans ist, gibt noch keinen Aufschluss darüber, in welchem Körper das Individuum steckt oder ob es in der Lage ist, eine Reproduktionsfunktion zu erfüllen.

Vielmehr sind trans-Menschen Individuen, bei denen die praktische Identität nicht mit dem sozial anerkannten, normativen Prinzip der Vereinheitlichung der sozialen Positionen und Rollen übereinstimmt. Folglich ist es überzeugender, das dritte Geschlecht durch die praktische Identität, die sexuelle Orientierung oder durch die Wünsche der trans-Menschen zu definieren, wie zum Beispiel: *X ist genau dann trans, wenn X ein soziales Individuum ist und [trans] besetzt. X besetzt genau dann [trans], wenn sich X als trans identifiziert.*

In diesem Zusammenhang kann jedoch angenommen werden, dass Witt diese Definition des dritten Geschlechts kategorisch von den binären Geschlechtern unterscheiden würde und entsprechend dem dritten Geschlecht nicht den gleichen Stellenwert wie Mann/Frau zuschreiben würde. Denn weder die soziale Anerkennung spielt eine definierende Rolle, noch wird das dritte Geschlecht durch die sozial vermittelte Reproduktionsfunktion charakterisiert. Folglich wäre es laut Witts Terminologie kategorisch falsch, von einem dritten Geschlecht zu sprechen.

6. Zweiteilige Auffassung von „Frau“

Unter der Annahme, dass trans-Frauen unter den Begriff „Frau“ fallen, schlage ich eine Erweiterung von Witts Definition vor und somit den Einbezug der Geschlechtsidentifikation: Mittels einer zweiteiligen Auffassung des Begriffs „Frau“ werden trans-Frauen per Definition weder exkludiert noch marginalisiert.

6.1 Konzept 1: [Frau] als soziale Position

Für die Definition des ersten Konzepts stütze ich mich auf die bereits diskutierte Auffassung von Geschlecht von Witt. Geschlecht lässt sich folglich insofern definieren, als dass es sich um eine soziale Position handelt, die jedem sozialen Individuum essentiell ist und alle anderen sozialen Positionen und Rollen determiniert, die das soziale Individuum besetzt. Demzufolge ist es die soziale Anerkennung, die festlegt, ob das soziale Individuum eine Frau oder ein Mann ist.

Geschlecht als soziale Position: X besetzt genau dann die soziale Position [Frau], wenn X ein Mensch ist und wenn die sozial vermittelte weibliche Reproduktionsfunktion „empfangen und gebären“ sozial anerkannt wird.

Die Definition des ersten Konzepts weicht von Witts Darstellung ab, weil hier weder davon ausgegangen wird, dass das Geschlecht eine essentielle Eigenschaft ist, die jedem sozialen Individuum zukommt, noch davon, dass das Geschlecht eine extrinsische Eigenschaft sozialer Individuen ist. Die ontologische Unterscheidung zwischen Mensch, Person und sozialem Individuum, die Witt trifft, sowie die Tatsache, dass das Geschlecht nur sozialen Individuen essentiell ist, spielt für meine Darstellung von „Geschlecht als soziale Position“ keine bedeutende Rolle. Einerseits wäre es verwirrend, von zwei unterschiedlichen Konzepten auszugehen, bei denen Geschlecht als Eigenschaft unterschiedlichen ontologischen Entitäten zugeschrieben wird. Andererseits handelt es sich bei der Frage, inwiefern die Begriffe „soziales Individuum“, „Person“ und „Mensch“ miteinander verbunden sind, und wo genau sich das Selbst verorten lässt, um ein philosophisches Problem, das ein ganz anderes Diskussi-

onsfeld eröffnet und entsprechend hier nicht weiter diskutiert werden kann.

Die notwendigen und hinreichenden Bedingungen, die erfüllt sein müssen, sodass nach dem ersten Konzept X eine Frau ist, sind demzufolge lediglich, dass X ein Mensch ist und als Besetzerin der sozialen Position [Frau] sozial anerkannt wird. Die soziale Anerkennung ist genau dann gewährleistet, wenn die körperlichen Eigenschaften darauf hindeuten, dass die sozial vermittelte weiblichen Reproduktionsfunktion *empfangen und gebären* vom Individuum erfüllt werden kann. In diesem Zusammenhang lässt sich sagen, dass dieses Konzept die metaphysische Frage, was eine Frau ist, aus der Perspektive der dominanten Geschlechter-Ideologie behandelt und demgemäss im Kontext dieser Ideologie einzubetten ist. Geschlecht ist eine soziale Position, die auf der sozialen Anerkennung körperlicher Eigenschaften in Bezug auf die sozial vermittelte Reproduktionsfunktion basiert.

6.2 Konzept 2: „Frau" als weibliche Geschlechtsidentität

Während das erste Konzept die soziale Realität einfängt und diejenigen Menschen als Frauen klassifiziert, die in der Gesellschaft als Frauen sozial anerkannt werden, wird laut dem zweiten Konzept mit dem Begriff „Frau" auf die Selbstidentifikation mit einer Geschlechtsidentität referiert. Inwiefern sich die Geschlechtsidentität charakterisieren lässt, ist eine komplexe Frage. Dennoch erachte ich die Auffassung als überzeugend, die Haslanger und Jenkins vertreten: Ihnen zufolge ist die Geschlechtsidentität als innere Karte aufzufassen, die ein Individuum durch die soziale und materielle Realität lenkt (Jenkins 2016, vgl. Haslanger 2005).

Haslanger beschreibt dazu eine autobiographische Analogie: Ihre persönliche Rassenidentität habe sich verändert, als sie ihre beiden schwarzen Kinder adoptierte. Obwohl sie sich nicht als schwarz identifiziere, beeinflusse und verändere die soziale und materielle Realität, d.h. die Hautfarbe ihrer Kinder sowie deren Erfahrungen und Reaktionen, ihre Identifikation als weisse Amerikanerin (vgl. Haslanger 2002, 31–35).[8] Jenkins definiert die Geschlechtsidentität wie folgt:

> *"S has a gender identity of X if S's internal map is formed to guide someone classed as a member of X gender through the social or material realities that are, in that context, characteristic of X as a class." (Jenkins 2016, 410)*

Die Definition des zweiten Konzepts wurzelt auf dieser Erläuterung, wobei die Klassifizierung mit einem Geschlecht mit der Besetzung der sozialen Position ausgetauscht wird, sodass die Terminologie mit der Definition des ersten Konzepts übereinstimmt. Entsprechend definiere ich die Geschlechtsidentität wie folgt:

> *X hat genau dann eine weibliche Geschlechtsidentität, wenn X's innere Karte insofern geformt ist, als dass sie einen Menschen, der die soziale Position [Frau] besetzt, durch die sozialen und materiellen Realitäten lenkt, die in diesem Kontext charakteristisch für die soziale Position [Frau] sind.*

X ist demnach laut dem zweiten Konzept genau dann eine Frau, wenn sich X als Frau identifiziert, d.h. wenn X eine weibliche Geschlechtsidentität hat. Die weibliche Geschlechtsidentität ist mit der sozialen Position [Frau] verknüpft, jedoch nicht dadurch determiniert. Dies deshalb, weil die innere Karte, die das Individuum lenkt und somit die Geschlechtsidentität konstituiert, an die soziale Position [Frau] und somit an die dominante Ideologie, d.h. an die zweigeschlechtliche sowie heterosexuelle Matrix gebunden ist. Demzufolge ist die Bedeutung der Geschlechtsidentität nicht fixiert, sondern kontextabhängig. Die Identifikation mit einem Geschlecht ist folglich ein Produkt aus den Interaktionen im dominanten System.

Es sei jedoch betont, dass die Identifikation als Frau keine notwendige Akzeptanz internalisierter Normen der Femininität impliziert. Vielmehr lässt sich die Interaktion zwischen dem Individuum, das sich als Frau identifiziert und den femininen Normen insofern charakterisieren, als dass das Individuum die femininen Normen als relevant betrachtet, wobei unterschiedliche Formen der Reaktion möglich sind. Das Gefühl einer Frau, dass keinen BH zu tragen unattraktiv und nicht feminin ist, prägt die Geschlechtsidentität dieser Frau aus dem Grund, weil das Gefühl auf die feminine Norm, wie der Körper einer Frau aussehen sollte, referiert. Trägt diese Frau entsprechend einen BH, vermeidet sie eine Verletzung einer dominanten Norm femininer Erscheinung. Eine andere Frau, die sich bewusst entscheidet, keinen BH zu tragen, verletzt zwar die dominanten femininen Normen, steht aber dennoch in Interaktion mit den dominanten Normen, da sie weiss, dass Frauen normkonform einen BH tragen sollten.

Die Geschlechtsidentität beinhaltet demzufolge ein subjektives und ein objektives Element: Das subjektive Element betrifft die Beurteilung des Subjekts, welche

8 Konferenz "Contextualist Approaches to the Metaphysics of Gender" 15.–16. Juni 2019 Humboldt Universität zu Berlin.

Normen für das Subjekt relevant sind. Die Geschlechtsidentität ist zudem objektiv im Sinne, als dass eine Verbindung zwischen dem subjektiven Element und den Normen besteht, die mit der relevanten sozialen Position assoziiert werden. Demgemäss stimme ich Jenkins zu, wenn sie davon ausgeht, dass ein Individuum nur dann eine Geschlechtsidentität gemäss diesem Konzept besitzt, wenn das Individuum ein minimales Verständnis davon hat, was die Normen bezüglich der dominanten Ideologie sind. Das minimale Verständnis beinhaltet beispielsweise die Kenntnis darüber, dass die Pronomen *sie/ihr* für Menschen verwendet werden, die die soziale Position [Frau] besetzen (Jenkins 2016). Entsprechend lässt sich die Geschlechtsidentität als Reaktion auf die sozialen Normen beschreiben, die mit den sozialen Rollen verbunden sind, die das Geschlecht als soziale Position konstituieren. Die Auffassung, dass die Geschlechtsidentität durch die dominante Ideologie geformt und geprägt wird, lässt sich insofern kritisieren, als dass die soziale Realität repressive, marginalisierende und diskriminierende Strukturen beinhaltet. Diesbezüglich kann argumentiert werden, dass die weibliche Geschlechtsidentität folglich ebenso inhärent unterdrückt ist.

Trotz der Tatsache, dass einige Normen des femininen Verhaltens die Unterdrückung der Frauen bestätigen, lässt sich die innere Karte jedoch auch insofern interpretieren, als dass resistentes und emanzipatorisches Verhalten gegenüber den femininen Normen möglich ist. Entsprechend beinhaltet die Identifikation als Frau kein inhärent repressives Moment. Trotz der Tatsache, dass die Frau, die keinen BH trägt, die Norm der femininen Erscheinung verletzt, ist es zumindest vorstellbar, dass sich dieser Akt positiv auf die Frau auswirkt, da sie sich dadurch emanzipiert fühlt und mit den femininen Normen der dominanten Ideologie bricht. Sich als Frau zu identifizieren, kann demzufolge in unterschiedlichen Kontexten für unterschiedliche Menschen unterschiedliche Dinge bedeuten. Denn die soziale und materielle Realität referiert auf ganz unterschiedliche Aspekte der eigenen verkörperten Existenz.

Entsprechend ist es kein Widerspruch, wenn cis-Frauen und trans-Frauen unterschiedliche Faktoren angeben, die ihre Geschlechtsidentität konstituieren: So begreift beispielsweise eine trans-Frau das Bedürfnis, dass sie in der Gesellschaft als Frau wahrgenommen werden möchte, als ihre weibliche Geschlechtsidentität; eine andere trans-Frau charakterisiert ihre weibliche Geschlechtsidentität als Wunsch, weibliche Genitalien zu haben.[9]

7. Deskriptive Definition von „Frau“ als zweiteilige Auffassung

Der Begriff „Geschlecht“ lässt sich laut der zweiteiligen Auffassung einerseits als soziale Position sowie andererseits als Geschlechtsidentität definieren. Diesbezüglich stellt sich die Frage, inwiefern die zwei unterschiedlichen Konzepte verbunden sind, sodass daraus eine deskriptive Definition des Begriffs „Frau“ formuliert werden kann. Mein Vorhaben unterscheidet sich demnach grundlegend vom Projekt von Jenkins, die zwar gleichermassen davon ausgeht, dass sich „Frau“ nur durch eine zweiteilige Auffassung definieren lässt. Sie verfolgt jedoch schlussendlich ein amelioratives Projekt und rechtfertigt die zweiteilige Auffassung in Hinblick auf feministische Ziele. Meines Erachtens lässt sich jedoch der Begriff „Frau“ nicht bloss ameliorativ, sondern deskriptiv als zweiteilige Auffassung definieren. Demgemäss sind die zwei Konzepte des Begriffs „Frau“ nicht im Hinblick auf ein praktisches Ziel oder eine politische und moralische Agenda formuliert, sondern fungieren als deskriptive Analyse auf die sokratische Frage, was eine Frau ist.

Die kritische Auseinandersetzung mit unterschiedlichen deskriptiven Projekten sowie das Prinzip der Transinklusion liess mich diesbezüglich folgern, dass sich die vielfältigen Aspekte der Metaphysik des Geschlechts durch ein einziges Konzept nicht erfassen lassen. Erst durch die Kombination der Konzepte „Frau“ als soziale Position sowie „Frau“ als Geschlechtsidentität lassen sich notwendige und hinreichende Bedingungen angeben, die festlegen, wann ein Individuum eine Frau ist. Demzufolge ist ein Individuum genau dann eine Frau, wenn die notwendigen und hinreichenden Bedingungen von mindestens einem Konzept erfüllt sind:

> Deskriptive Definition des Begriffs Frau: *X ist genau dann eine Frau, wenn X ein Mensch ist und die soziale Position [Frau] besetzt oder eine weibliche Geschlechtsidentität hat.*[10]

Die zwei Konzepte, auf die der Begriff „Frau“ referiert, konstituieren die Definition des Begriffs „Frau“, wobei beide Konzepte gleich relevant sind und den gleichen theoretischen Status haben. Die zwei Konzepte stehen weder in einem Spannungsverhältnis, noch sind sie als

9 Dabei handelt es sich um eine Simplifizierung. Die Geschlechtsidentität ist eine komplexe Angelegenheit, die in den meisten Fällen aus unterschiedlichen Elementen zusammengesetzt ist: Elemente, die den eigenen Körper, den Geist sowie soziale Interaktionen betreffen.

10 Dabei handelt es sich um ein inklusives „oder“: X ist demgemäss auch dann eine Frau, wenn X ein Mensch ist und die soziale Position [Frau] besetzt und eine weibliche Geschlechtsidentität hat.

separierte dichotomische Phänomene zu betrachten. Vielmehr handelt es sich bei den zwei Konzepten um unterschiedliche Aspekte eines mehrdeutigen Begriffs, wobei die deskriptive Definition die Konzepte im System vereint.

Die deskriptive Definition erfüllt das Prinzip der Transinklusion, indem die Identifikation als Frau von gleicher Relevanz ist, um als Frau zu gelten, wie die soziale Positionierung in der Gesellschaft mit zusammenhängender sozialer Anerkennung. Entsprechend sind die notwendigen und hinreichenden Bedingungen, um als Frau zu gelten, auch bei einer trans-Frau erfüllt, die in der Gesellschaft nicht als Frau anerkannt wird, sich jedoch als Frau identifiziert. Ein weiterer Vorteil meiner Definition liegt darin, dass nicht nur trans-Frauen unter den Begriff der Frau fallen, sondern auch Frauen, die biologisch weibliche Sexmerkmale besitzen und sich als Frau identifizieren, jedoch aufgrund ihrer maskulinen äusseren Erscheinung sozial nicht als Frauen anerkannt werden (Watson 2016, 247). Cis-Frauen erfüllen womöglich in den meisten Fällen die notwendigen und hinreichenden Bedingungen beider Konzepte. Nicht zuletzt geht die zweiteilige Auffassung mit unseren Intuitionen einher, da die existierenden femininen Normen, die sozial vermittelte Reproduktionsfunktion, körperliche Eigenschaften sowie die soziale Realität und die Unterdrückungsstrukturen in die Definition miteinbezogen werden, jedoch keine notwendigen Bedingungen sind, um als Frau zu gelten.

Literatur

- Aristoteles. 1959. *Über die Zeugung der Geschöpfe. Die Lehrschriften* Bd. 8,3: Paderborn: Ferdinand Schöningh.
- Berger, Peter und Luckmann, Thomas. 1966. *The social construction of reality.* Garden City NY: Anchor Books.
- Bettcher, Talia Mae. 2009. "Trans Identities and First-Person Authority." In *You've Changed: Sex Reassignment and Personal Identity*, edited by Laurie Shrage, 98–120. Oxford, New York: Oxford University Press.
- ——— 2012. "Trans Women and the Meaning of 'Woman'." In *The Philosophy of Sex: Contemporary Readings*, 6. edition, edited by Power, Nicholas, Raja Halwani, and Alan Soble. 233–250. Lanham, MD: Rowman and Littlefield Publishers.
- ——— 2014. "Feminist Perspectives on Trans Issues." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Sept 26, 2009.
- De Beauvoir, Simone. 1949. *Le Deuxième Sexe.* Paris: Gallimard.
- Haslanger, Sally. 2002. "Gender and Race: (What) are they? (What) do we want them to be?" *Nous* 34 (1): 31–55.
- ——— 2005. "You mixed? Racial Identity without racial Biology." In *Adoption Matters. Philosophical and Feminist Essays*, edited by Haslanger, Sally and Charlotte Witt, 265–289. Ithaca: Cornell University Press.
- ——— 2007. *Metaphysics of Gender. Standford Encyclopedia of Philosophy.*
- Jeffreys, Sheila. 2014. *Gender Hurts: A Feminist Analysis of the Politics of Transgenderism.* New York: Routledge.
- Jenkins, Katharine. 2016. "Amelioration and Inclusion: Gender Identity and the Concept of Woman." *Ethics* 126 (2): 394–421.
- Mikkola, Mari. 2008. *Feminist Perspectives on Sex and Gender.* In *The Standford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Mai 12, 2008.
- ——— 2010. "Ontological Commitments, Sex and Gender." In *Feminist Metaphysics Explorations in the Ontology of Sex, Gender and the Self*, edited by Witt, Charlotte, 67–84. New York: Springer.
- Saul, Jennifer. 2012. "Politically significant terms and philosophy of language: Methodological issues." In *Out from the shadows: Analytic feminist contributions to traditional philosophy,* edited by Crasnow, Sharon and Anita Superson. Oxford: Oxford University Press.
- Watson, Lori. 2016. "The Woman Question." *Transgender Studies Quarterly* 3 (1–2): 246–253.
- Witt, Charlotte. 1995. "Anti-Essentialism in Feminist Theory." *Philosophical Topics: Feminist Perspectives on Language, Knowledge and Reality* 23 (2): 321–44.
- ——— 2011. *The Metaphysics of Gender.* Oxford: University Press.

Rahel Wehrlin (26) hat Philosophie und Geschichte studiert und schreibt an einer interdisziplinären Dissertation mit dem Titel „Die schweizerische Pornographie-Subkultur 1988-2020. Aneignung als queerfeministische Praxis und Wissensproduktion" an der Universität Bern. Sie interessiert sich für Queer Theory, Kritische Theorie, Antike Philosophie und für den Deutschen Idealismus.