Die Fakten sprechen lassen: Werte und Wertfreiheit in gesellschaftlich relevanter Forschung

MARKUS DRESSEL

DIE FAKTEN SPRECHEN LASSEN:
WERTE UND WERTFREIHEIT
IN GESELLSCHAFTLICH RELEVANTER FORSCHUNG

ZUSAMMENFASSUNG Welche Rolle spielen Werturteile in wissenschaftlicher Erkenntnis – und welche Implikationen ergeben sich hieraus für das Verhältnis von Wissenschaft und Gesellschaft? Die vorliegende Arbeit betrachtet diese Fragen aus wissenschaftsphilosophischer Perspektive. In Teil I wird die Diskussionslage zum Ideal wertfreier Wissenschaft dargestellt. Teil II präsentiert eine Begriffsanalyse des Wertfreiheitsideals, wobei insbesondere dessen Teilthesen, Einschränkungen und Interpretationsmöglichkeiten betrachtet werden. In Teil III wird einer der wichtigsten philosophischen Einwände gegen Wertfreiheit diskutiert: das Argument des induktiven Risikos. Teil IV behandelt den Ansatz Philip Kitchers zur Integration von Bürgerinnen und Bürgern in Forschungsprozesse. Im Zentrum steht dabei das Konzept der idealen Deliberation. In Teil V wird ein theoretisch informiertes Verständnis von Wissenschafts-Gesellschafts-Modellen vorgelegt und dessen Nutzung als Reflexions-Tool in realen Wissenschafts-Gesellschafts-Interaktionen diskutiert.

SCHLAGWORTE Werte, Wertfreiheit, Wissenschaft, Wissenschafts-Gesellschafts-Verhältnis

ABSTRACT What role do value-judgements play in science – and what are the consequences for the relation between science and society? This work addresses these issues from a philosophy of science perspective. Part I presents the current and classic debate on the ideal of value-free science. Part II is a conceptual analysis of the value-free ideal, with special consideration of this ideal's sub-claims, its restrictions and conceivable interpretations. Part III discusses one of the most prominent philosophical critiques of value-freedom: the argument from inductive risk. Part IV is devoted to Philip Kitcher and his call for an intensified participation of citizens in science. In particular, this part focusses on the concept of an ideal deliberation between citizens and scientists. Part V develops a theoretical understanding of science-society models and discusses how this understanding can be used as a reflexive tool in actual science-society interactions.

KEY WORDS values, value-freedom, science, science-society relation

Die Fakten sprechen lassen: Werte und Wertfreiheit in gesellschaftlich relevanter Forschung

MARKUS DRESSEL

Zitierhinweis:
Dressel, M. (2023). Die Fakten sprechen lassen: Werte und Wertfreiheit in gesellschaftlich relevanter Forschung (Dissertation). Gottfried Wilhelm Leibniz Universität Hannover. https://doi.org/10.15488/13199

Gestaltung: Jan Henrik Arnold
Schrift: Stempel Garamond LT Pro

Für Yvonne.

Inhalt

III. INDUCTIVE RISK: DOES IT REALLY REFUTE VALUE-FREEDOM?

IV. CITIZEN PARTICIPATION IN KITCHER'S WELL-ORDERED SCIENCE: WHAT KIND OF IDEAL IS DELIBERATION?

V. MODELS OF SCIENCE AND SOCIETY: TRANSCENDING THE ANTAGONISM

Vorwort

Alte philosophische Fragen sind ein zweischneidiges Schwert. Einerseits üben sie, gerade weil sie trotz ihrer langen Geschichte weiterhin unbeantwortet scheinen, eine starke Anziehung aus. Andererseits bergen sie besondere Gefahren: Die Diskussionslage ist bereits so ausdifferenziert, dass das Thema in eine Fülle schwer zu überblickender Einzelaspekte zerfällt; gleichzeitig steigt das Risiko, nichts Substanzielles mehr beitragen zu können; und schließlich haben solche Themen eine gewisse Neigung zur Lager- und Schulenbildung. Und doch sind diese Fragen besonders reizvoll. Denn sie eröffnen die Möglichkeit, an fundamentalen, unser Denken und Handeln in vielfacher Hinsicht beeinflussenden Grundsatzthemen zu arbeiten. Letztlich entspringt ihr Reiz somit jener Motivation, die Hannah Arendt als Antrieb ihres philosophischen Wirkens beschrieben hat: *Ich will verstehen*.

Das Thema dieser Arbeit ist die Rolle von Werten in der wissenschaftlichen Erkenntnis. Zweifellos handelt es sich dabei um ein Thema genau dieser Art – die diskutierten Fragen sind alt, die Argumente komplex und die Fronlinien klar gezogen. Dass ich mich dennoch mit diesem Thema beschäftige, hat vor allem zwei Gründe. Zum einen ist die Weise, wie Wissenschaft betrieben wird, von größter Wichtigkeit für die Herausforderungen unserer Zeit. Beispiele hierfür gibt es viele, wobei die COVID-19-Pandemie und der Klimawandel vielleicht die eindrücklichsten sind. Je nachdem, wie Wissenschaftlerinnen und Wissenschaftler mit Werturteilen im Forschungsprozess umgehen, werden ihre Beiträge hierzu andere sein. Die Folgen der zunächst eher theoretisch erscheinenden Frage nach Werten in der Wissenschaft sind somit keineswegs nur akademischer Natur.

Zum anderen scheint mir, dass es trotz des beeindruckenden Literaturstands noch etwas zu diesem Thema beizutragen gibt. So habe ich die wichtigsten Teilaspekte etwas ausführlicher und, so zumindest meine Hoffnung, begrifflich schärfer als üblich rekonstruiert. Einige der hierauf beruhenden Argumente gehen über das bloße Wiederholen des Altbekannten hinaus und sind, wie ich ebenfalls hoffe, tatsächlich von philosophischer Bedeutung. Außerdem habe ich versucht, die Einseitigkeiten zu vermieden, die ich in Teilen der Debatte zu erkennen glaube. Nach Jahrzehnten der Verteidigung des Wertfreiheitsideals ist das Pendel der philosophischen Diskussion nun in die andere Richtung geschwungen, so dass es beim Lesen neuerer Beiträge zuweilen unverständlich erscheint, wie vernünftige Menschen jemals das Ideal wertfreier Wissenschaft vertreten konnten.

Obwohl ich viele der wertfreiheitskritischen Argumente teile, scheint mir die Endgültigkeit, mit der das Scheitern des Wertfreiheitsideals – und mit ihm bestimmter Modelle des Wissenschafts-Gesellschafts-Verhältnisses – mitunter verkündet wird, ein wenig extrem. Ein differenzierter Blick auf die alte Frage nach der Wertfreiheit oder -beladenheit von Forschung lohnt sich daher nicht nur, er tut auch Not.

Die vorliegende Arbeit ist zugleich das Ergebnis eines mehrjährigen philosophischen Projektes und einer persönlichen Reise. Dass ich auf dieser Reise von ebenso kompetenten wie wohlwollenden Unterstützern begleitet wurde, ist ein unschätzbarer Wert. Torsten Wilholt hat mir mit unzähligen klugen Kommentaren beim Schärfen meiner Positionen geholfen und mich selbst dann zum Ausarbeiten dieser Positionen ermutigt, wenn sie seinen eigenen widersprachen. Das ist keineswegs selbstverständlich. Ebenso wenig selbstverständlich ist seine Unterstützung bei den vielen kleineren und größeren Rückschlägen, die ein solches Projekt mit sich bringt. Hierfür bedanke ich mich herzlich. Dietmar Hübner hat meine Arbeit als scharfsinniger, aufmerksamer und überaus verlässlicher Diskussionspartner begleitet. In den Gesprächen mit ihm habe ich nicht nur viel gelernt, sie haben mich auch vor einer Reihe misslicher Fehler bewahrt. Dafür bin ich sehr dankbar.

Hermann Held hat mir die Fortführung dieses Projektes in Situationen ermöglicht, in denen nicht klar war, ob und wann es letztendlich abgeschlossen sein wird. In der Tat stand dieses Projekt mehr als einmal vor dem Abbruch – dass es nicht dazu gekommen ist, verdanke ich zu nicht geringen Teilen ihm. Hinzu kommen die vielen anregenden und vertrauensvollen Gespräche, die ich in den vergangenen Jahren mit ihm führen durfte. Auch hierfür bin ich sehr dankbar. Danken möchte ich weiterhin den Mitgliedern des DFG-Graduiertenkollegs „Integrating Ethics and Epistemology of Scientific Research“. Ohne den intellektuellen Input und die praktische Unterstützung des Kollegs wäre dieses Projekt so nicht möglich gewesen. Ähnliches gilt für die vielen Personen, die mir bei einzelnen Teilen der vorliegenden Arbeit geholfen haben (sie werden am Ende der entsprechenden Teile erwähnt). Keine dieser Personen hat jedoch dieses Projekt so sehr mitgetragen – und mit erlitten – wie meine geliebte Partnerin und Verlobte Yvonne Ehnert. Ihr ist diese Arbeit gewidmet: Danke für deine liebevolle Unterstützung, deine Teilnahme und deine Geduld.

I. EINLEITUNG: VIEFALT UND RELEVANZ DER WERTFREIHEITSDEBATTE

ZUSAMMENFASSUNG Welche Rolle spielen ethische, soziale und politische Werturteile in der Wissenschaft – und welche Rolle sollten sie spielen? Im ersten Teil dieser Arbeit gebe ich einen Überblick über die philosophische Wertfreiheitsdebatte. Ziel ist es, die thematische Vielfalt und die Relevanz dieser Debatte herauszuarbeiten. Ich beginne mit ihrem historischen Ausgangspunkt: dem Ideal wertfreier Wissenschaft. Ich skizziere verschiedene Begründungsmuster für das Wertfreiheitsideal und beleuchte die hinter dem Ideal stehende Motivation. Den wertfreiheitsstützenden Argumenten stelle ich eine Reihe wertfreiheitskritischer Argumente gegenüber. Weiterhin zähle ich die wichtigsten Strategien auf, die Kritikerinnen und Kritiker des Wertfreiheitsideals für den Umgang mit Werturteilen vorgeschlagen haben. Die jeweiligen Argumente und Ansätze stelle ich dabei knapp dar, ohne sie an dieser Stelle zu bewerten. Eine Ausnahme bildet das *Sein-Sollen-Argument*. Dieses aus meiner Sicht bedeutende Argument behandle ich etwas ausführlicher, weil es in der Wertfreiheitsdebatte häufig in anderer Form oder nur oberflächlich diskutiert wird. Außerdem bereite ich damit eine Auseinandersetzung in einem späteren Teil dieser Arbeit vor. Auch die übrigen Ausführungen dienen dazu, auf Diskussionen in den folgenden Teilen hinzuführen. Ein Überblick über die Einzelteile und die Struktur der Gesamtarbeit folgt am Ende des ersten Teils.

1. Was auf dem Spiel steht: Wertfreiheit aus Sicht ihrer Vertreterinnen und Vertreter

1.1 WOZU WERTFREIHEIT I: ARTEN VON BEGRÜNDUNGEN

Seit Max Weber (1904/1988; 1917/1988; 1919/1988) zu Beginn des letzten Jahrhunderts die Trennung von wissenschaftlicher Analyse und außerwissenschaftlichen Wertungen forderte, ist eine ebenso umfangreiche wie kontroverse Debatte über das Für und Wider von wissenschaftlicher Wertfreiheit entbrannt (z.B. Fleck 1935/1980; Reichenbach, 1938/1961; Rudner, 1953; Jeffrey, 1956; Kuhn, 1962; Hempel, 1965; Popper, 1974; Kuhn, 1977; McMullin, 1982; Longino, 1990; Harding, 1995; Lacey, 1999; Douglas, 2000; Kourany; 2003; Giere, 2003; Kitcher, 2011; Betz, 2013; Bright, 2018; Holman & Wilholt, 2022). Dass diese Debatte nach über einhundert Jahren weiterhin geführt wird, zeigt, dass es um etwas geht. Tatsächlich reichen die Implikationen der Wertfreiheitsdebatte weit über akademische Problemstellungen hinaus und betreffen letztlich die Frage, welches Verhältnis von Wissenschaft und Gesellschaft angemessen ist: Wie kann sichergestellt werden, dass Wissenschaft gesellschaftsrelevante Probleme aufnimmt und auf legitime Weise in gesellschaftliche Entscheidungen einbringt? Und wie kann umgekehrt gewährleistet werden, dass gesellschaftliche Einflüsse den Forschungsprozess nicht unangemessen verzerren? Angesichts der Bedeutung, die Wissenschaft für Themen wie Klimapolitik, technologischen Fortschritt oder öffentliche Gesundheit hat, sind diese Fragen nicht nur theoretischer Natur – was auf dem Spiel steht, ist unser Verständnis von *guter Wissenschaft* und *guten wissenschaftsgestützten Entscheidungen*.

Diese Gegenüberstellung – die Integrität des Erkenntnisprozesses einerseits und die Legitimität wissenschaftlicher Gesellschaftsberatung andererseits – weist bereits auf die Quellen jenes Ideals hin, das von Max Weber erstmals in seiner modernen Form vertreten und später vielfach übernommen oder kritisiert wurde: das Ideal wertfreier Wissenschaft. Die zentrale Forderung dieses Ideals lautet:

> daß der Forscher und Darsteller die Feststellung empirischer Tatsachen […] und *seine* praktisch wertende, d. h. diese Tatsachen […] als erfreulich oder unerfreulich *beurteilende*, in diesem Sinn: ‚bewertende' Stellungnahme unbedingt *auseinanderhalten* solle (Weber, 1917/1988, S. 500, Hervorh. i. Orig.).

Der Ausgangspunkt der Wertfreiheitsdebatte liegt somit in der von Weber und seinen Nachfolgerinnen und Nachfolgern vertretenen These, dass Wissenschaft – insofern sie „Feststellung empirischer Tatsachen" ist – frei von wissenschaftsfremden Werturteilen sein soll. Ich werde den genauen Inhalt und mögliche Interpretationen des Wertfreiheitsideals später ausführlich diskutieren (Teil II und III). Um die Relevanz der Debatte zu verstehen, muss aber zunächst einmal die hinter dem Wertfreiheitsideal stehende *Motivation* geklärt werden (Proctor, 1991; Holman & Wilholt, 2022). Hier kommen die beiden erwähnten Quellen ins Spiel, nämlich einerseits *epistemische*, also auf den Erkenntnisprozess bezogene Begründungsmuster und andererseits *nicht-epistemische*, also auf außerwissenschaftlichen Motiven beruhende Begründungen. Obwohl diese Diskussionsstränge oft ineinander übergehen und sich gegenseitig bedingen, eröffnet die Unterscheidung ein systematisches Verständnis des *Wozu* des Wertfreiheitsideals: Im ersten Fall geht es um die Sorge, dass Werturteile die wissenschaftliche Qualität des Forschungsprozesses untergraben und so wichtige epistemische Güter wie Wahrheit oder logische Schlüssigkeit beschädigen; im zweiten Fall bezieht sich die Sorge auf die Beschädigung ethischer und politischer Güter, etwa die Autonomie der demokratischen Willensbildung.

1.2 WOZU WERTFREIHEIT II: EPISTEMISCHE BEGRÜNDUNGSMUSTER

Die Gruppe der epistemischen Begründungen beinhaltet mehrere zusammenhängende, im Detail aber verschiedene Argumente. Das *Heterogenitätsargument* geht davon aus, dass Wissenschaft wesentlich durch das Ziel der Wahrheitsfindung definiert ist und dass wissenschaftsfremde Werturteile nichts zu diesem Ziel beitragen. Der „Bestätigungsgrad von Tatsachenaussagen", so das Argument, „ist unabhängig davon, ob sie ethischen Wunschvorstellungen entsprechen" (Schurz, 2011, S. 180). Ähnlich betont bereits Weber, dass der ethische Wert einer Aussage nichts zu deren Wahrheit oder Falschheit beisteuert, „weil es sich da nun einmal um heterogene Probleme handelt" (Weber, 1917/1988, S, 500). Das *Bias-Argument* geht von derselben Annahme aus, behauptet aber zusätzlich, dass Werturteile zu Fehlern führen. So schrieb bereits Francis Bacon, auf den dieses Argument zurückgeht (Proctor, 1991): „Der menschliche Verstand gleicht einem Spiegel mit unebener Fläche für die Strahlen der Gegenstände, welcher seine Natur mit der der letzteren vermengt, sie entstellt und verunreinigt" (1620/2017, S. 44). Diese „Verunreinigung" kann nach dem Bias-Argument jedoch vermieden werden, wenn sich Wissenschaftlerinnen und Wissenschaftler der ethischen, politischen oder persönlichen Wertung enthalten. Das *Sein-Sollen-Argument* ähnelt dem Bias-Argument. Anders als

dieses interpretiert es wertbeladene Wissenschaft jedoch nicht als empirisch inadäquat, sondern als logisch fehlschlüssig (z.B. Haack, 2003, S. 13). Grundlage hierfür ist eine „Schlussbarriere", die das Ableiten von deskriptiven Konklusionen aus normativen Prämissen verbietet. Auf dieses Argument gehe ich gleich noch ausführlicher ein.

Das *Unbegründbarkeitsargument* beruht ebenfalls auf der Trennung unterschiedlicher Aussagetypen, konzentriert sich aber auf das Wesen von Werturteilen anstatt auf deren Effekte. Es behauptet, dass Werturteile keinen Wahrheitswert annehmen und somit nicht wissenschaftlich diskutiert werden können (Weber, 1919/1988; Reichenbach, 1938/1961). Daher sollen, so das Argument, Wissenschaftlerinnen und Wissenschaftler normative Annahmen weder im Forschungsprozess noch in der öffentlichen Kommunikation von Forschungsergebnissen verwenden. Das *Politisierungsargument* behauptet, dass Wissenschaft „die für sie spezifische Funktion" (Luhmann, 1990, S. 296) der Wahrheitsfindung nur aufgrund ihrer „autonomen Geschlossenheit" (ebd.) erfüllen kann. Hiernach importieren Werturteile, da sie meist kontrovers sind, die systemfremde Logik des politischen Streits in die Wissenschaft, wodurch diese dysfunktional wird (Weingart, 2001; Strohschneider, 2014). Das *Fehlsteuerungsargument* geht in eine ähnliche Richtung. Nach diesem Argument ist Wissenschaft am effektivsten, wenn sie von Fragen der Praxisanwendung entlastet ist; orientiert sie sich hingegen an außerwissenschaftlichen Werten und Bedürfnisse, allokiert sie Ressourcen für wissenschaftlich irrelevante und häufig unlösbare Probleme (Kuhn, 1962, S. 37; Polanyi, 1962).

Nicht alle diese Diskussionsstränge sind gleichermaßen bedeutend für die philosophische Debatte, und manche von ihnen beziehen sich auf Aspekte des Forschungsprozesses, die über das Ideal wertfreier Wissenschaft im engeren Sinne hinausgehen. Dies wird in Teil II deutlicher werden. Dennoch weisen die genannten Begründungsmuster auf die Vielfalt der Gefahren hin, die wertbeladene Wissenschaft aus Sicht der Befürworterinnen und Befürworter von Wertfreiheit haben kann. Unabhängig davon, welche dieser Argumente im Einzelnen vertreten werden, verfolgen sie jedoch dasselbe Ziel: Schaden vom Erkenntnisprozess abzuwenden. Die Sorge um Wissenschaft ist somit ein zentrales Motiv in der Wertfreiheitsdebatte – allerdings, wie wir gleich sehen werden, nicht das einzige.

1.3 WOZU WERTFREIHEIT III: NICHT-EPISTEMISCHE BEGRÜNDUNGSMUSTER

Im Gegensatz zu den epistemischen Begründungsmustern betrachten die nicht-epistemischen Wissenschaft nicht um ihrer selbst willen, sondern wegen der Auswirkungen, die sie auf andere normativ relevante Güter hat. Der gemeinsame Ausgangspunkt dieser nicht-epistemischen Argumente ist die Wichtigkeit von Wissenschaft für praktische Entscheidungen. So sind Forschungsergebnisse unverzichtbar für die politische Regulation von Substanzen und Medikamenten, aber auch für die Bewältigung gesellschaftlicher Herausforderungen wie des Klimawandels. Auch auf individueller Ebene, etwa bei Konsumentscheidungen, können Forschungsergebnisse handlungsleitend sein. Und schließlich kann Wissenschaft unser Denken und Handeln auf einer allgemeineren Ebene beeinflussen, wie sich unter anderem in gesellschaftlichen Diskursen über Geschlechterverhältnisse oder Tierethik zeigt. Nicht-epistemische Argumente für Wertfreiheit versuchen, den Nutzen von Wissenschaft für diese Praxiszusammenhänge zu stärken und Schaden von ihnen abzuwenden.

Am deutlichsten zeigt sich dies am *pragmatischen Argument*. Nach dieser Überlegung hat wertbeladene Wissenschaft typischerweise einen geringen Praxisnutzen als wertfreie. Da Problemlösungen ein Verständnis der Faktenlage voraussetzen, „werden wissenschaftliche Hypothesen ohne empirische Basis nur sehr eingeschränkt brauchbar sein" (Koertge, 2013, S. 250). Das oben erwähnte Bias-Argument, wonach Wertbeladenheit empirische Fehlurteile begünstigt, wird hier also um die Annahme erweitert, dass derartige Fehler die Praxistauglichkeit von Forschung untergraben. Einen etwas anderen Weg wählt das *Universalismusargument*. Hiernach können wertbeladene Forschungsergebnisse durchaus nützlich sein, jedoch nur für diejenigen Anwender, für deren Bedürfnisse sie produziert wurden. Wissenschaft soll aber, so das Argument, unabhängig von partikularen Nutzungsinteressen verwertbar sein (Merton, 1942/1973; Jeffrey, 1956) – insbesondere, weil sich diese Interessen ändern können, ohne dass sich deswegen unsere empirischen Überzeugungen ändern sollten (Koertge, 2013). Das *Glaubwürdigkeitsargument* beruht auf ähnlichen Annahmen, vertritt jedoch die zusätzliche Hypothese, dass die Öffentlichkeit wertbeladener Wissenschaft ihr Vertrauen entziehen wird. Zumindest für Fälle, in denen die verwendeten Werturteile denen der jeweiligen Bürgerinnen und Bürger widersprechen, ist dies auch tatsächlich empirisch belegt (Elliott et al., 2017). Das *kritische Argument* fügt dem noch eine Überlegung hinzu: Nur dann, wenn Wissenschaft keine Rücksicht auf das ethisch oder politisch Erwünschte nehmen muss, kann sie auf Missstände hinweisen und so der Gesamtgesellschaft nutzen (Schurz, 2013).

Nach diesem Argument hat Wertfreiheit somit eine politisch-emanzipatorische Dimension (ebd., S. 310).

Das *Präskriptionsargument* betrachtet Wissenschaft aus der Perspektive eines liberalen Demokratie- und Freiheitsverständnisses. Hiernach ist Wertfreiheit geboten, weil Wissenschaft nicht zu Werturteilen legitimiert ist: „As political decisions are informed by scientific findings, the valuefree ideal ensures […] that collective goals are determined by democratically legitimized institutions, and not by a handful of experts" (Betz, 2013, S. 207). Dies lässt sich auch auf individuelle Entscheidungen übertragen, wo Wertfreiheit nicht den demokratischen Prozess, sondern die Autonomie einzelner Wissenschaftsanwenderinnen und -anwender schützen soll (Weber, 1919/1988; Betz, 2013). Das *Fairnessargument* verfolgt eine ähnliche Motivation. Da Forschungsprozesse nicht frei zugänglich sind, bedeutet Wertbeladenheit nach dieser Überlegung eine unangemessene Bevorteilung der beteiligten Wissenschaftlerinnen und Wissenschaftler. Dies wird auch von Kritikerinnnen und Kritikern des Wertfreiheitsideals anerkannt: „to the extent that scientists make value judgments, there are concerns that their values will be undemocratically privileged" (Intemann, 2015, S. 218). Das *Ideologieargument* geht von demselben Demokratieverständnis aus, verknüpft dies aber mit einer Version des kritischen Arguments. Hiernach läuft wertbeladene Wissenschaft Gefahr, zum Instrument illiberaler oder gar totalitärer Politik zu werden (Merton, 1942/1973). Als Beispiele werden häufig die sogenannte „Deutsche Physik" der Nationalsozialisten oder die lyssenkoistische Biologie im stalinistischen Russland diskutiert (z.B. John, 2019). Um Derartiges bereits im Ansatz zu verhindern, dürfen außerwissenschaftliche Werturteile nach dem Ideologieargument generell, d.h. unabhängig von ihrem konkreten Inhalt, nicht zugelassen werden.

Wie man sehen kann, sind die genannten epistemischen und nicht-epistemischen Begründungen eng miteinander verknüpft, und zwar sowohl innerhalb der jeweiligen Gruppen als auch zwischen ihnen. Die Abgrenzung der einzelnen Diskussionsstränge ist daher ein Stück weit Ermessenssache. Gleichzeitig sind diese Diskussionen zweifellos komplexer, als es die skizzierten Argumente widerspiegeln. Ich werde auf einige dieser Argumente noch genauer eingehen (s. das folgende Kapitel sowie Teil III). Dennoch unterstreicht bereits der kursorische Überblick die Relevanz der durch das Wertfreiheitsideal adressierten Probleme. Dies zeigt sich nicht zuletzt daran, dass Kritikerinnen und Kritiker von Wertfreiheit viele dieser Bedenken explizit anerkennen (z.B. Douglas, 2009; Intemann, 2015; John, 2019; Holman & Wilholt, 2022). Fraglich ist daher nicht, ob die jeweiligen epistemischen und nicht-epistemischen Güter schützenswert sind, son-

dern ob Wertfreiheit hierfür zweckdienlich ist und, sollte dies nicht der Fall sein, durch welche alternativen Mittel sie ersetzt werden kann. In jedem Fall gilt aber, worauf Hilary Putnam einmal mit drastischen Worten hingewiesen hat: „the question as to what the differences are between ‚factual' judgments and ‚value' judgments is no ivory-tower issue. Matters of – literally – life and death may well be at stake" (2002, S. 2).

2. Begründung des Wertfreiheitsideals am Beispiel des Sein-Sollen-Arguments

2.1 INHALT UND AUFBAU DES ARGUMENTS

Um die Relevanz der Wertfreiheitsdiskussion weiter zu akzentuieren, möchte ich eines der aufgezählten Begründungsmuster ein wenig genauer betrachten: das *Sein-Sollen-Argument*. Dabei handelt es sich zwar nicht um eine „matter of life and death"; dennoch thematisiert das Argument ein Gut von fragloser Bedeutung: die Übereinstimmung von Wissenschaft mit logischen Prinzipien. Darüber hinaus ist diese Begründung auch deswegen interessant, weil sie (anders als etwa das Bias- oder das pragmatische Argument) ohne Annahmen über die Folgen von Wertbeladenheit auskommt. Das Sein-Sollen-Argument motiviert das Wertfreiheitsideal also selbst dann, wenn neben der Verletzung logischer Prinzipien keine weiteren Konsequenzen zu befürchten wären. Aus diesem Grund werde ich das Thema auch später noch einmal aufgreifen (Teil III). Da es dort jedoch mehr um die Frage geht, ob Wissenschaft wertbeladen und dennoch logisch unproblematisch sein kann, möchte ich das Argument hier zunächst als solches vorstellen. Eine mögliche Formulierung lautet[1]:

1. Schlüsse von deskriptiven auf normative Aussagen sind ungültig (*Kein-Sollen-aus-Sein*).
2. Wenn Schlüsse von deskriptiven auf normative Aussagen ungültig sind, dann sind auch Schlüsse von normativen auf deskriptive Aussagen ungültig (*Kein-Sein-aus-Sollen*).

1 Da ich hier stärker auf die begrifflichen Grundlagen des Arguments eingehen möchte (insbesondere die Prinzipien Kein-Sollen-aus-Sein und Kein-Sein-aus-Sollen), wähle ich eine andere Formulierung als in Teil III. Dort setzte ich diese Prinzipien schlicht voraus, ohne ihren Hintergrund zu beleuchten. Dies spiegelt sich auch in der Wahl der Terminologie („Sein-Sollen-Argument" statt „Argument from Wishful Thinking"). Im Kern geht es jedoch um dieselbe Problematik.

3. Wissenschaftliches Wissen muss als Menge rein deskriptiver Aussagen verstanden werden.
4. Daher dürfen Aussagen, wenn sie wissenschaftliches Wissen instanziieren können sollen, nicht aus normativen Aussagen erschlossen werden.

Wie sich zeigen wird, bedarf dieses Argument einiger Spezifikationen und Einschränkungen. Weiterhin impliziert das Argument nicht automatisch die Forderung, dass sich Wissenschaftlerinnen und Wissenschaftler aller wissenschaftsfremden Werturteile enthalten sollen – denn es ist ja denkbar, dass sie Pflichten *neben* der Wissensproduktion haben. Ich werde darauf noch zurückkommen. Bevor ich die Anwendung des Arguments auf Wissenschaft diskutiere, möchte ich jedoch den Ausgangspunkt des Arguments betrachten, nämlich das Prinzip *Kein-Sollen-aus-Sein*.

2.2 DER ERSTE SCHRITT: SCHLÜSSE VOM SEIN AUFS SOLLEN

Der erste Schritt des Arguments betrachtet das logische Verhältnis von *Ist*-Aussagen (etwa empirische Beobachtungen) und *Sollen*-Aussagen (etwa ethische Normen). Ziel dieses Schrittes ist die Etablierung einer „barrier to is-ought inferences" (Guevara, 2008, S. 46), die Schlüsse von ersteren auf letztere blockiert. Obwohl es sich dabei um die Gegenrichtung des in der Wertfreiheitsdebatte thematisierten Schlusstyps handelt, bietet sich diese Vorüberlegung wegen ihrer größeren Bekanntheit an. Der *locus classicus* hierfür ist David Hume (1740/1978). Hume kritisierte die Angewohnheit zeitgenössischer Autoren, aus empirischen Beobachtungen Aussagen abzuleiten, in denen ihm „anstatt der üblichen Verbindungen von Worten mit ‚*ist*' und ‚*ist nicht*' kein Satz mehr begegnet, in dem nicht ein ‚*sollte*' oder ‚*sollte nicht*' sich fände" (ebd., S. 211). Da derartige Sätze „eine neue Beziehung" (ebd.) ausdrücken, sei es „ganz unbegreiflich […] wie diese neue Beziehung zurückgeführt werden kann auf andere, die von ihr ganz verschieden sind" (ebd.). Damit legte Hume die Grundlage jenes logischen Prinzips, das heute oft als *Kein-Sollen-aus-Sein* bezeichnet wird.

Wie ist dieses Prinzip begründet? Eine Möglichkeit, Kein-Sollen-aus-Sein zu plausibilisieren, basiert auf dem sogenannten Konservationsprinzip der Logik. Nach diesem Grundsatz sind die Konklusionen eines gültigen Schlusses deswegen gültig, weil sie die Eigenschaft der Gültigkeit gewissermaßen von den Prämissen erben (Pidgen, 2016). Logik handelt demnach von notwendigen Beziehung: Konklusionen *müssen* gültig sein,

wenn es die Prämissen sind. Dies scheint aber nur dann zu funktionieren, wenn die Konklusionen ausschließlich Inhalte aufweisen, die bereits in den Prämissen enthalten sind. Denn ansonsten wäre es, wie Hume schreibt, „ganz unbegreiflich", woher die Eigenschaft der Gültigkeit in den prämissenfremden Teilen der Konklusion stammt. Da logische Schlüsse somit auf dem Grundsatz „you cannot get out what you haven't put in" (Pidgen, 2010, S. 217) beruhen, scheinen Schlüsse von deskriptiven auf normative Aussagen prinzipiell ungültig.

Es zeigt sich jedoch, dass Kein-Sollen-aus-Sein nur mithilfe weiterer Spezifikationen aufrechterhalten werden kann. Die erste Spezifikation besteht in der Erläuterung, dass die Schlussbarriere nur *direkte* Sein-Sollen-Schlüsse blockiert, also solche, die entweder gar keine normativen Prämissen enthalten („Harry Potter ist ein Zauberer; daher darf Harry Potter zaubern") oder die den normativen Gehalt der Konklusion ohne korrekte Verwendung der normativen Prämissen erzeugen („Harry Potter ist ein Zauberer; Zauberer dürfen Quidditch spielen; daher darf Harry Potter zaubern"). Offensichtlich gilt dies aber nicht für *indirekte* Schlüsse, in denen deskriptive Prämissen lediglich zur Konklusion beitragen, ohne deren normativen Gehalt zu generieren („Harry Potter ist ein Zauberer; Zauberer dürfen Quidditch spielen; daher darf Harry Potter Quidditch spielen"). Diese Spezifikation ist naheliegend und scheint nicht von besonderem philosophischen Interesse. Wie ich jedoch später zeigen werde (Teil III), ist die Direktheitsbedingung wichtig, denn sie eröffnet zentrale Möglichkeiten für einen logisch unproblematischen Umgang mit Werten im Forschungsprozess.

Eine zweite Spezifikation lässt sich aus einer klassischen Kritik von Arthur Prior (1960) ableiten. Für Prior ist Kein-Sollen-aus-Sein kein universelles Prinzip, da es Schlüsse gibt, die aus deskriptiven Prämissen normative Konklusionen erzeugen und dennoch gültig sind. Ein Beispiel lautet etwa: „There is no man over 20 feet high; therefore there is no man over 20 feet high who is allowed to sit in an ordinary chair" (Prior, 1960, S. 202). Im Gegensatz zu Prior sind jedoch die meisten Autorinnen und Autoren der Ansicht, dass derartige Beispiele das Prinzip nicht widerlegen, sondern lediglich einschränken: auf substanzielle oder „nicht-leere" (Pidgen, 2010) Schlüsse. Priors Schlüsse sind hingegen „leer" in dem Sinne, dass der normative Teil der Konklusion arbiträr hinzugefügt anstatt aus den Prämissen abgeleitet wird. Das Beispiel könnte daher ebenso lauten „Es existiert kein über 20 Fuß großer Mann; daher existiert kein über 20 Fuß großer Mann oder Harry Potter darf Quidditch spielen". Solche Schlüsse sind verpflichtungsirrelevant (Schurz, 1997), weil der normative Teil durch beliebige Propositionen – auch durch sein genaues Gegenteil – ersetzt werden kann. Ein beliebiges Werturteil ist jedoch gar kein Werturteil,

da hieraus kein *Sollen* folgt: „we can't advance our normative inquiry by considering [such] an argument" (Singer, 2015, S. 200). Da es sich nicht um echte Werturteile handelt, sind die „leeren" Schlüsse Priors nicht Gegenstand des Sein-Sollen-Arguments.

Die dritte Spezifikation ergibt aus einer Kritik von John Searle und Alasdair MacIntyre. Auch hier geht es um Schlüsse, die scheinbar Kein-Sollen-aus-Sein widerlegen, etwa: „Jones promised to pay Smith five dollars […] [therefore:] Jones ought to pay Smith five dollars" (Searle, 1964, S. 44; ähnlich MacIntyre, 1981/2013, S. 68). Offensichtlich verstehen solche Einwände eine Prämisse nur dann als normativ, wenn der normative Gehalt explizit, d.h. in Form deontischer Operatoren („soll" etc.) oder Prädikate („gut" etc.) vorliegt. Da „P verspricht x" keine solchen Begriffe enthält, scheint es, als würde eine Sollen-Aussage aus einer Ist-Aussage erschlossen. Dies ist aber nicht plausibel. Denn offenbar können normative Gehalte auch in impliziter Form vorliegen, nämlich in der Semantik der in der Prämisse verwendeten Begriffe (Pidgen, 2016). Eben dies scheint in Searles Beispiel der Fall. Wer nicht weiß, dass Versprechen Realisierungspflichten erzeugen, hat schlicht den Begriff „Versprechen" nicht verstanden. Searles Beispiel beruht damit letztlich auf dem Missverständnis, dass Kein-Sollen-aus-Sein derartige Fälle von „analytic entailment" (Pidgen, 2016, S. 406) ausschließt. Ein sinnvolles Verständnis von Kein-Sollen-aus-Sein bezieht sich hingegen nur auf Fälle, in denen normative Gehalte nicht einfach aus einer semantischen Analyse der Prämissen abgeleitet werden.

2.3 DER ZWEITE SCHRITT: SCHLÜSSE VOM SOLLEN AUFS SEIN

Nach den obigen Überlegungen ist Kein-Sollen-aus-Sein ein plausibles Prinzip, wenn es auf *direkte*, *nicht-leere* und *nicht semantisch implizierte* Schlüsse beschränkt wird. Wie steht es aber mit dem zweiten Schritt des Sein-Sollen-Arguments, nach dem die Schlussbarriere auch in die Gegenrichtung gilt? Interessanterweise wird das entsprechende Prinzip, *Kein-Sein-aus-Sollen*, in der Wertfreiheitsdebatte eher als *erkenntnistheoretisches* anstatt als *logisches* Prinzip diskutiert. Im Mittelpunkt steht somit häufig das Überraschungspotenzial der Welt gegenüber unseren Präferenzen, seltener aber das Problem der logischen Schlüssigkeit. Einschlägige Beiträge finden sich aber in anderen Zusammenhängen, etwa der Philosophie der Emotionen oder der Sprachphilosophie. So argumentieren etwa Justin D'Arms und Daniel Jacobson (2000) in einem vieldiskutierten Artikel, dass die moralische Fragwürdigkeit einer Emotion nichts über deren faktisches Passungsverhältnis mit der Welt aussagt. Daher, so die Autoren, kann etwa ein Witz gleichzeitig unmoralisch und witzig sein. In ähnlicher Weise kritisierte Tom

Campbell (1970) die Angewohnheit der damaligen Sprachphilosophie, Annahmen über tatsächliches Sprechverhalten aus normativen Prinzipien herzuleiten. Beide Artikel machen ausdrücklich Gebrauch von Kein-Sein-aus-Sollen.

In der Wertfreiheitsdebatte wird das Prinzip zwar häufig erwähnt (wobei das Problem meist mit dem Begriff des *Wunschdenkens* umschrieben wird, s. auch Teil III dieser Arbeit); jedoch geschieht dies häufig *en passant* (z.B. Haack, 2003; Koertge, 2013) und mitunter in einer Weise, die dem Thema nicht gerecht zu werden scheint. So sind Sollen-Sein-Schlüsse etwa für Matthew Brown (2013) unproblematisch, solange die normativen Prämissen Gegenstand ergebnisoffener Diskurse bleiben. Elizabeth Anderson (2004) betrachtet Werturteile gar als empirisch falsifizierbare Hypothesen. In beiden Fällen soll die Schlussbarriere überwunden werden, indem Werturteile als fallibel betrachtet werden. Das ist jedoch wenig überzeugend. Selbst wenn diese Vorschläge die erkenntnistheoretische Frage, wie Wissenschaft wertbeladen sein und dennoch unwillkommene Ergebnisse liefern kann, beantworten würden (was als solches bereits zweifelhaft ist), so bliebe doch die logische Frage, wie normative Prämissen deskriptive Konklusionen implizieren können. Denn dieses Problem wird offensichtlich nicht dadurch gelöst, dass andere normative Gehalte in die Prämissen eingesetzt werden – der Schluss „Harry Potter soll Quidditch spielen; daher spielt Harry Potter Quidditch" ist eben auch dann unzulässig, wenn die Prämisse durch ein anderes Werturteil ersetzt wird.

Ist es stattdessen möglich, Kein-Sein-aus-Sollen mit ähnlichen Argumenten anzugreifen, wie ich sie im Zusammenhang mit Kein-Sollen-aus-Sein diskutiert habe? Auch dies ist wenig aussichtsreich. Wie gezeigt, beruhen diese Argumente entweder auf nicht-substantiellen bzw. „leeren" Schlüssen oder auf impliziten normativen Gehalten auf Prämissenseite. Dasselbe trifft nun auch für Kein-Sein-aus-Sollen zu: In einem „leeren" Sollen-Sein-Schluss wie „Harry Potter darf Quidditch spielen; daher darf Harry Potter Quidditch spielen oder Harry Potter spielt Quidditch" ergibt sich der deskriptive Teil der Konklusion nicht aus der Prämisse, sondern ist arbiträr hinzugefügt. Ebenso kann ist „Harry Potter darf Quidditch spielen; daher ist Harry Potter ein Zauberer" nur dann logisch unproblematisch, wenn die Semantik von „Quidditch" eine Beschränkung auf Zauberer enthält. Dann ist die Prämissenmenge aber nicht mehr rein normativ, da sie den deskriptiven Satz „Wer Quidditch spielen darf, ist ein Zauberer" implizit enthält. Folglich ist es auch im Fall von Kein-Sein-aus-Sollen nicht sinnvoll, das Prinzip auf *leere* oder *semantisch implizierte* Schlüsse anzuwenden.

Dies bedeutet nun aber nicht, dass der „inferential gap between ‚ought' and ‚is'" (Guevara, 2008, S. 46) jedwede Präsenz normativer Elemente in der Prämissenmenge ausschließt. Denn auch die dritte Einschränkung von Kein-Sollen-aus-Sein, die Direktheitsbedingung, trifft ebenso in der Gegenrichtung zu. Solange normative Prämissen die deskriptive Konklusion nicht *erzeugen*, sondern lediglich dazu *beitragen*, können Sollen-Sein-Schlüsse offenbar unproblematisch sein. So entspringt etwa in „Harry Potter darf Quidditch spielen; wer Quidditch spielen darf, ist ein Zauberer; daher ist Harry Potter ein Zauberer" das „ist" in der Konklusion nicht aus der normativen, sondern aus der deskriptiven Prämisse. Obwohl dieser Tatbestand in logischer Hinsicht unspektakulär ist, ist er für die Wertfreiheitsdebatte von großem Interesse. So basiert einer der wichtigsten Vorschläge zur Verwendung von Werturteilen in der Wissenschaft auf eben dieser Direktheitsbedingung. Ich werde diesen vor allem von Heather Douglas (2000, 2009) vertretenen Vorschlag in Teil III dieser Arbeit diskutieren. Dabei wird sich zeigen, dass Wissenschaft wertbeladen und dennoch logisch unproblematisch sein kann, wenn Werturteile lediglich als „Tiebreaker" zwischen wissenschaftlichen Entscheidungsoptionen mit gleichem oder ähnlichem epistemischen Erwartungsnutzen fungieren. Ansonsten bleibt es jedoch dabei, dass Sollen-Sein-Schlüsse, wenn sie nicht indirekt, leer oder semantisch impliziert sind, „ganz unbegreiflich" (Hume, 1740/1978, S. 211) sind.

2.4 DER DRITTE SCHRITT: WISSENSCHAFT ALS „STUDY OF FACTS"

Im dritten Schritt des Sein-Sollen-Arguments wird behauptet, dass Forschungsergebnisse als Mengen deskriptiver Aussagen verstanden werden müssen. Dies ist wichtig, da ansonsten die in den vorherigen Schritten etablierte Schlussbarriere irrelevant für die wissenschaftliche Erkenntnis wäre. Dabei muss einem Missverständnis vorgebeugt werden: die Anwendung von Kein-Sein-aus-Sollen auf Wissenschaft setzt nicht voraus, dass diese dem oben vorgeführten Schema logischen Schließens folgt. Obwohl Forschungsprozesse auch formale (z.B. mathematische) Elemente enthalten, sind Forschungsergebnisse meist nicht logisch erzwungen, sondern höchstens wohlbegründet. Dasselbe trifft aber auf die meisten Lebensbereiche zu, ohne dass wir der Meinung sind, logische Fehlschlüsse seien deswegen gestattet. Vertreterinnen und Vertreter des Sein-Sollen-Arguments müssen daher nicht behaupten, dass Wissenschaft gewissermaßen angewandte Logik sei, sondern lediglich, dass Wissenschaft logischen Prinzipien nicht *widersprechen* soll. Die Voraussetzung hierfür ist jedoch, dass das Ziel von Wissenschaft im Generieren deskriptiver Aussagen besteht.

In der Tat ist diese dritte Behauptung des Sein-Sollen-Arguments äußerst plausibel. Nach Ansicht vieler Autorinnen und Autoren soll Wissenschaft uns nicht darüber informieren, was der Fall sein soll, sondern darüber, was der Fall ist: „Science, by definition, is the study of the way the world is, the study of facts" (Bělohrad, 2011, S. 265). Dies kann zum einen methodologisch begründet werden, denn die für Wissenschaft spezifischen Verfahren sind offensichtlich nicht geeignet, normatives Wissen *aus sich heraus* zu generieren. Da hierfür andere Verfahren (ethische Reflektion, politische Deliberation etc.) zur Verfügung stehen, besteht schlicht kein Grund, diese Verfahren durch die ungeeigneten wissenschaftlichen Verfahren zu ersetzen. Zum anderen scheint es gerade die Deskriptivität wissenschaftlichen Wissens zu sein, die Wissenschaft wertvoll für praktische Zwecke macht. Dies wird auch von Kritikerinnen und Kritikern des Wertfreiheitsideals betont. So argumentiert etwa Heather Douglas: „We desire scientific expertise because we want predictively reliable accounts of the world on which to base our decisions, that is, accounts such that if we act on that basis, we are likely to get the predicted results" (2008, S. 2). Ähnlich begründet Philip Kitcher sein Konzept der „epistemischen Arbeitsteilung" mit dem Nutzen, den Bürgerinnen und Bürger aus deskriptiven Informationen ziehen: „the deference to experts is appropriate because those experts help [the citizens] overcome the limitations of their knowledge, and thus to formulate and pursue their freely chosen projects more effectively" (2011, S. 21; s. auch Teil IV dieser Arbeit). Beide Aspekte, der Vorhersageerfolg wissenschaftlicher Informationen und die Effektivität der hierauf beruhenden Handlungen, wären aber offenbar gefährdet, wenn Wissenschaft uns nicht darüber informieren würde, was tatsächlich (oder wahrscheinlich) der Fall ist. Insofern scheint der Praxisnutzen wissenschaftlichen Wissens auf seiner Deskriptivität zu beruhen.

Somit scheint auch der dritte Schritt des Sein-Sollen-Arguments überzeugend. Bedeutet dies, dass Forschungsprozesse automatisch fehlschlüssig sind, wenn sie ethische oder politische Werturteile beinhalten? Wie ich bereits erwähnt habe, ist dies nicht notwendig der Fall: die logische Problematik kann vermieden werden, indem die Direktheitsbedingung von Kein-Sein-aus-Sollen umgangen wird, d.h. indem normative Prämissen nicht als propositionale Quelle deskriptiver Konklusionen fungieren. Ich werde dieses Thema in Teil III etwas ausführlicher diskutieren.

Weiterhin schließt das Sein-Sollen-Argument nicht aus, dass Wissenschaftlerinnen und Wissenschaftler Pflichten *neben* der Produktion wissenschaftlichen Wissens haben. Ein klassisches Beispiel hierfür sind ethische Methodenbeschränkungen, wie sie etwa in klinischer Forschung verwendet werden. Die Verantwortung gegenüber den beteiligten

Probandinnen und Probanden ist nicht durch das Ziel der Wahrheitsfindung begründet, sondern durch Werturteile ethischer Art. Dennoch entstehen hierdurch keine logischen Probleme, denn diese Werturteile sind nicht in dem Sinne Teil der Prämissenmenge, dass aus ihnen deskriptive Gehalte erschlossen werden. Vielmehr beschränken sie die Prämissenmenge, indem sie das Spektrum der verfügbaren Daten limitieren. Ähnliches ließe sich auch für anderen Werturteile zeigen, etwa das von Kevin Elliott (2011) vorgeschlagene „no-passing-the-buck-principle". Nach diesem Prinzip haben Wissenschaftlerinnen und Wissenschaftler die Pflicht, der Gesellschaft auch dann wissenschaftliche Informationen zur Verfügung zu stellen, wenn diese mit gewissen Unsicherheiten behaftet sind. Auch hier werden jedoch keine deskriptiven Konklusionen aus normativen Prämissen erzeugt – zumindest solange nicht, wie die Entscheidung, ein Forschungsergebnis trotz bestehender Unsicherheiten an die Öffentlichkeit zu kommunizieren, als *ethische* oder *politische,* nicht aber als *wissenschaftliche* Entscheidung betrachtet wird. Somit stellt das Sein-Sollen-Argument zwar eine starke Begründung des Wertfreiheitsideals dar; diese Begründung schließt jedoch nur bestimmte Arten von Wertbeladenheit aus. Solange deskriptive Aussagen nicht unmittelbar aus normativen erschlossen werden, können Werturteile auch nach dem Sein-Sollen-Argument eine akzeptable Rolle im Forschungsprozess spielen.

3. Wertfreiheit aus Sicht ihrer Kritikerinnen und Kritiker

3.1 POPULÄRE EINWÄNDE I: KRITIK DES SOLLENSANSPRUCHS

Wie ich bereits angedeutet habe, hat die philosophische Debatte eine große Zahl wertfreiheitskritischer Argumente hervorgebracht. Ähnlich wie die diskutierten Begründungsmuster für Wertfreiheit gehen auch die kritischen Argumente häufig ineinander über und bedingen sich gegenseitig. Um einen geordneten Überblick zu erhalten, ist es jedoch sinnvoll, zwischen *normativen* und *deskriptiven* Kritiken zu unterscheiden (die dahinterstehende Systematik erörtere ich in Teil II). Normative Argumente betreffen den Sollensanspruch des Wertfreiheitsideals, also die Frage, ob Wertfreiheit tatsächlich erstrebenswert ist. Im Zentrum stehen dabei häufig die Folgen, die das Streben nach Wertfreiheit für ethische und politische, teilweise aber auch für epistemische Güter

hat. Deskriptive Argumente beziehen sich auf die Möglichkeit von Wertfreiheit, wobei mit der Möglichkeitskritik die (implizite oder explizite) Behauptung verknüpft ist, dass unerreichbare Ziele schlechte Ziele sind (diesen Zusammenhang diskutiere ich ausführlich in den Teilen II und IV).

Die Gruppe der normativen Kritiken ist weniger umfangreich als die der deskriptiven. Der Diskussionsstrang der *axiologischen Kritik*[2] bezieht sich auf den Nutzen ethischer und politischer Werturteile für den Forschungsprozess. Anstatt diese als Gefahr für epistemische Güter zu betrachten, behauptet der Einwand das genaue Gegenteil: „values can be good for science" (Longino, 2004, S. 127). Hiernach können Werturteile etwa helfen, Verzerrungen in traditioneller Forschung zu korrigieren (Longino, 2004; Hicks, 2014). Diese Argumentation tritt häufig gemeinsam mit der *emanzipatorischen Kritik* auf. In einer frühen Variante dieser Kritik ging etwa Jürgen Habermas von einem „emanzipatorischen Erkenntnisinteresse" (1965/2013, S. 67) aus, das Machtverhältnisse aufzudecken und einen „herrschaftsfreien Dialog aller mit allen" (1965/2013, S. 71) zu realisieren versucht. In der neueren Diskussion vertritt unter anderem Janet Kourany eine ähnliche Position: „scientists [should] include only specific social values in science, namely the ones that meet the needs of society" (2008, S. 95), wobei Kourany insbesondere an feministische Werte denkt. Das *Maskierungsargument* hängt eng mit der axiologischen und der emanzipatorischen Kritik zusammen. Anders als diese betont es jedoch weniger den Nutzen von Werturteilen als die schädliche Wirkung des Wertfreiheitsideals. Die Orientierung an Wertfreiheit, so das Argument, „can have the dangerous consequence of masking the influence of contextual factors" (Biddle, 2013, S. 131). Anstatt Werturteile zu verdecken, sollen sie nach diesem Argument besser anerkannt und kritisch diskutiert werden.

Ein als *aims approach* bekanntes Argument (Elliott & McKaughan, 2014; Intemann, 2015) behauptet, dass Wissenschaft nicht nur dem Ziel der Wahrheitsfindung, sondern auch gesellschaftlichen Zielen verpflichtet ist. Folglich sei etwa die Priorisierung der Anwendbarkeit eines Forschungsergebnisses ein legitimes Werturteil, das sogar epistemische Ziele wie empirische Adäquatheit übertrumpfen kann[3]. Eine besonders umfäng-

2 Diese Bezeichnung ist angelehnt an Holman & Wilholt (2022). Während es dort jedoch um „Demarkationsstrategien" geht, also um die Abgrenzung legitimer und illegitimer Werturteilsgebräuche, geht es mir hier um Typen werturteilskritischer Argumente.

3 Je nach Interpretation kann dieser Ansatz so verstanden werden, dass er im Widerspruch zu der im letzten Kapitel diskutierten These des Sein-Sollen-Arguments steht, nach der Forschungsergebnisse Mengen deskriptiver

liche Diskussion ist schließlich mit dem *Argument des induktiven Risikos* verbunden (z.B. Douglas, 2000; Wilholt, 2009; Steel, 2016; Biddle & Kukla, 2017). Dieses erstmals in den 1950er Jahren in seiner modernen Form formulierte Argument (Rudner, 1953) gilt bis heute als "[o]ne of the most important reasons for thinking that non-epistemic values can play a legitimate role in scientific reasoning" (Elliott & Steel, 2017, S. 6). Da Forschungsergebnisse immer mit einer gewissen Unsicherheit einhergehen, müssen Wissenschaftlerinnen und Wissenschaftler nach diesem Argument festlegen, wieviel Sicherheit für die Anerkennung einer Hypothese notwendig ist. Diese Entscheidung kann jedoch Folgen für ethische und politische Güter haben, etwa wenn sich wissenschaftliche Aussagen über die Toxizität von Stoffen als falsch herausstellen. Nach der normativen Lesart des Arguments *sollen* Wissenschaftlerinnen und Wissenschaftler Werturteile fällen, indem sie diese Konsequenzen in die Festlegung akzeptabler Fehlerwahrscheinlichkeiten einbeziehen (eine ausführliche Diskussion dieses Arguments folgt in Teil III).

3.2 POPULÄRE EINWÄNDE II: KRITIK DES MÖGLICHKEITSANSPRUCHS

Neben den normativen Kritiken existiert eine größere Zahl deskriptiver, also auf die Möglichkeit von Wertfreiheit gerichteter Kritiken. Das *Argument des induktiven Risikos* findet sich auch in dieser Gruppe, hier allerdings in seiner deskriptiven Lesart. Danach ist es nicht nur unmoralisch, die außerwissenschaftlichen Folgen einer Hypothesenbewertung zu ignorieren – es ist schlichtweg *unmöglich*, da ansonsten unklar wäre, woher der Maßstab zur Festlegung akzeptabler Fehlerwahrscheinlichkeiten stammt (Rudner, 1953; Wilholt, 2009; Biddle & Winsberg, 2010; Winsberg, 2012). Das *Unterdeterminationsargument* ist eng damit verwandt. Nach diesem auf W. V. O. Quine (1951) und Pierre Duhem (1906/1954) zurückgehenden Argument kann eine gegebene Menge von Daten von mehr als einer Theorie erklärt werden, und zwar entweder *immer* („globale Unterdetermination") oder zumindest vor dem Hintergrund des *aktuellen* Forschungsstands („transiente Unterdetermination") (Biddle, 2013). Die Theoriewahl muss daher, so das Argument, durch außerwissenschaftliche Präferenzen entschieden werden (Longino, 1990; Kourany, 2003).

Aussagen sind. Wie in diesem Kapitel angedeutet, halte ich ein solches Wissenschaftsverständnis für irreführend. Meiner Ansicht nach handelt es sich bei den im aims approach diskutierten Zielen nicht um Ziele neben der Wahrheitsfindung, sondern um Spezifikationen der Art von Wahrheiten, die in bestimmten gesellschaftlichen Kontexten angestrebt werden. Diese Auffassung liegt auch meinen Argumenten in Teil III zugrunde.

Ein weiterer Diskussionsstrang lässt sich unter der Bezeichnung der *Abgrenzungskritiken* zusammenfassen. Eine dieser Kritiken bezieht sich auf die Unterscheidbarkeit von *epistemischen und nicht-epistemischen Werten* (ich diskutiere die beiden Werturteilsarten in Teil II). Hiernach sind wissenschaftliche Präferenzen, etwa für einfache oder generalisierende Theorien, untrennbar mit gesellschaftlichen Werturteilen verbunden. Dem Argument zufolge ist es daher unmöglich, wissenschaftliche Entscheidungen ausschließlich aufgrund wissenschaftlicher Präferenzen herbeizuführen (Longino, 2004). In einer Variante dieser Kritik sind epistemische und nicht-epistemische Werte deswegen untrennbar, weil sie für sich genommen nicht hinreichend informativ sind: „epistemic values […] are individually imprecise and can conflict with each other" (Wilholt, 2009, S. 97; s. auch Kuhn, 1977). Andere Abgrenzungskritiken beziehen sich auf die Unterscheidbarkeit wertfreier und wertbeladener *Kontexte* (Hoyningen-Huene, 2006) (für eine ausführlichere Diskussion siehe Teil II). Eine mit Thomas S. Kuhn (1962) verknüpfte Kritik argumentiert, dass sich der Kontext der *Rechtfertigung* einer Theorie nicht von ihrer häufig durch Idiosynkrasien und Zufälle gekennzeichneten *Entdeckungsgeschichte* trennen lässt (s. auch Fleck, 1935/1980). Eine weitere, in der Tradition John Deweys (1939) stehende Abgrenzungskritik weist die Trennung von Rechtfertigungs- und Anwendungskontexten zurück. Hiernach sind wissenschaftliche Theorien *Mittel* zur Erreichung praktischer *Zwecke*. Da Mittel und Zwecke jedoch untrennbar miteinander verknüpft seien, ist Wissenschaft nach diesem Argument notwendig wertbeladen: „policy objectives and the [scientific] means are highly interdependent and cannot be evaluated separately" (Edenhofer & Kowarsch, 2015).

Einen etwas anderen Schwerpunkt legt die *semantische Kritik* (Putnam, 2002; Dupré, 2013). Diesem Argument zufolge müssen viele wissenschaftliche Begriffe als „thick concepts" (Williams, 1985), also als hybride Ausdrücke mit deskriptivem und normativem Gehalt verstanden werden[4]. John Dupré illustriert dies anhand eines drastischen Beispiels: „Wer meint, dass er die Ursache von Vergewaltigung untersucht, aber […] dies ohne Vorurteil darüber tut, ob Vergewaltigung etwas Gutes oder Schlechtes sei, unterliegt einer schweren Konfusion: Er versteht gar nicht, was er angeblich untersucht" (2013, S. 264). Das *Framingargument* geht in eine ähnliche Richtung. Es behauptet, dass

4 Dieses Argument steht nur dann im Widerspruch mit meinen Ausführungen zum Sein-Sollen-Argument, wenn normative und deskriptive Gehalte *niemals* voneinander getrennt werden können. Selbst Vertreterinnen und Vertreter der semantischen Kritik scheinen dies jedoch nicht in dieser Radikalität zu vertreten. In begrenztem Umfang sind „thick concepts" jedoch mit dem Sein-Sollen-Argument kompatibel, da ein Sollen-Sein-Schluss, dessen Prämissenmenge derartige Ausdrücke enthält, als *semantisch impliziert* verstanden werden kann.

die Lösung eines wissenschaftlichen Problems durch seine Formulierung prästrukturiert wird. Insbesondere bei gesellschaftsrelevanten Themen seien Problemformulierungen jedoch von sozialen Perspektiven beeinflusst, weswegen Problemlösungen diesem Argument zufolge abhängig von sozialen Werturteilen sind. Dieses Problem- und Wissenschaftsverständnis wird auch mit den Begriffen „wicked problems" (Rittel & Webber, 1973) oder „post-normal science" (Funtowicz & Ravetz, 1993) umschrieben. Damit verwandt ist die aus der soziologischen Wissenschaftsforschung stammende *empirische Kritik*. Diese Kritik beruft sich auf eine Vielzahl von Studien, in denen Wissenschaft als *de facto* wertbeladen charakterisiert wurde (Latour & Woolgar, 1979; Collins, 1985; Gibbons et al., 1994). Da Wertfreiheit nicht empirisch nachgewiesen werden könne, so die Kritik, muss sie als „figment of philosophical imagination" (Hedgecoe, 2004, S. 131) abelehnt werden.

4. Strategien für den Umgang mit wertbeladener Wissenschaft

4.1 STRATEGIEN I: DIE FORDERUNG NACH WERTURTEILSTRANSPARENZ

Die vorangegangenen Überlegungen verdeutlichen die zentrale Herausforderung der Wertfreiheitsdebatte: Einerseits spricht eine Reihe plausibler, auch von Kritikerinnen und Kritikern anerkannter Gründe für das Ideal wertfreier Wissenschaft, andererseits spricht eine mindestens ebenso große Zahl guter Gründe gegen dieses Ideal. Es stellt sich daher nicht nur die Frage, ob das Wertfreiheitsideal – möglicherweise in modifizierter Form – aufrechterhalten werden kann (ich diskutiere diese Frage in Teil III); es muss auch geklärt werden, welche Alternativen die Gegnerinnen und Gegner von Wertfreiheit anzubieten haben. Denn sollten ihre Kritiken zutreffen, dann löst die Zurückweisung des Wertfreiheitsideals zwar die in diesen Kritiken angesprochenen Probleme. Gleichzeitig entstehen jedoch neue Probleme. So muss etwa gezeigt werden, wie Wissenschaft wertbeladen und dennoch nicht-präskriptiv, also kompatibel mit demokratischen Prinzipien und der Autonomie einzelner Wissenschaftsnutzerinnen und -nutzer sein kann (s. Teil III und Teil IV). Hieran zeigt sich abermals die Relevanz der Wertfreiheitsdebatte, denn die Zurückweisung des Wertfreiheitsideals hat Implikationen für die Frage, in welchem Verhältnis Wissenschaft und Gesellschaft grundsätzlich stehen können und sol-

len. Ich diskutiere das Thema der Wissenschafts-Gesellschafts-Modelle in Teil V dieser Arbeit. Da diese Modelle jedoch vielfältige Aspekte neben der Wertfreiheitsproblematik beinhalten, ist es sinnvoll, die Strategien zunächst als solche zu betrachten.

Eine interessante Beobachtung ist dabei, dass viele dieser Strategien einen gemeinsamen Ausgangspunkt haben: die Forderung nach Transparenz. Danach müssen Werturteile zunächst explizit gemacht werden, bevor ein sinnvoller Umgang mit ihnen gefunden werden kann. Diese Forderung ist nicht nur allgegenwärtig, sie wurde auch bereits früh in der Wertfreiheitsdebatte erhoben. So war Werturteilstransparenz etwa für Max Weber (1919/1988) eine Minimalbedingung der „Kathederwertung", also der Verwendung von Werturteilen in der akademischen Lehre. Auch für den Forschungsprozess selbst wird Ähnliches seit langem diskutiert. So argumentierte etwa Richard Rudner: „objectivity for science lies at least in becoming precise about what value judgments are being and might have been made in a given inquiry" (1953, S. 6).

Die Transparenzforderung bietet jedoch ihrerseits Anlass zu Fragen. Zum einen ist nicht klar, ob Werturteilstransparenz tatsächlich möglich ist. Zum anderen ist fraglich, ob Transparenz, wenn sie möglich ist, bereits hinreichend für einen unproblematischen Umgang mit Werturteilen ist. Was die Möglichkeit von Transparenz betrifft, so hat Eric Winsberg die These vertreten, dass Werturteile häufig in den Tiefen von Forschungsprozessen – ihren „nooks and crannies" (2012, S. 132) – verborgen sind. Ein Grund hierfür ist die Vielzahl der beteiligten Wissenschaftlerinnen und Wissenschaftler sowie die Unübersichtlichkeit der wissenschaftlichen Arbeitsteilung. Winsberg illustriert dies anhand komplexer Klimamodelle. Diese werden häufig über viele Jahre von unterschiedlichen Forschungsgruppen entwickelt, wobei sich der Beitrag einzelner Beteiligter jeweils auf Teilaspekte beschränkt. Wenn Werturteile in einer solchen „massively distributed collaboration" (Winsberg et al. 2014, S. 16) gefällt werden, dann lassen sie sich im Nachhinein kaum rekonstruieren oder gar durch alternative Werturteile ersetzen: „it becomes terribly hard to ask for climate science that reflects 'better' values" (Winsberg, 2012, S. 132).

Ein generelles Argument gegen Transparenz lässt sich hieraus jedoch nicht ableiten. Denn erstens ist nicht jeder Forschungsprozess „massively distributed". Zwar basiert Wissenschaft nahezu immer auf vorangegangener und ermöglicht ihrerseits nachfolgende Wissenschaft (Wilholt, 2013); dennoch ist die beschriebene Unübersichtlichkeit, zumindest in der Radikalität, ein spezifisches Merkmal wissenschaftlicher Großprojekte. Zweitens können Wissenschaftlerinnen und Wissenschaftler ihre *eigenen* Werturteile selbst dann transparent machen, wenn ihnen die Werturteile *anderer* verborgen bleiben.

Und drittens beinhaltet die Wertfreiheitsdebatte eine Vielfalt von Kandidaten für Werturteile, die keineswegs opak sind. Hierzu zählen etwa die Festlegung von Signifikanzlevels (Douglas, 2000), die Entscheidung für bestimmte Tiermodelle (Wilholt, 2009) oder die Wahl ökonomischer Diskontraten (Frisch, 2013). Wären diese potenziellen Werturteile tatsächlich unzugänglich, könnten sie nicht in dieser Weise diskutiert werden. Selbst wenn manche Werturteile in den „nooks and crannies" der wissenschaftlichen Arbeitsteilung verborgen bleiben, spricht dies somit nicht generell gegen Werturteilstransparenz.

Bedeutet dies, dass Transparenz, insoweit sie möglich ist, auch hinreichend ist? In der Tat scheinen einige Autorinnen und Autoren eine solche These zu vertreten (z.B. Schneider, 1997; Stanton, 2010; Elliott & Resnik, 2014). So argumentieren etwa Kevin Elliott und Daniel McKaughan: „if those engaged in assessing a hypothesis or theory are clear enough about their goals and the criteria […], those who disagree about those goals or criteria can 'backtrack' and adopt their own alternative assessments and conclusions" (2014, S. 15-16). In ähnlichler Weise fordern die Autoren eines für das Britische Außenministerium verfassten Berichts: „Any valuation of the risks of climate change will involve subjective judgments […]. Such judgments should be made transparently, so that they may be publicly debated" (King et al., 2015, S. 10).

Wie ich jedoch später argumentieren werde (Teil III), ist Transparenz nur ein schwacher Schutz gegen einige der erwähnten Argumente für das Wertfreiheitsideal (Steel, 2017), insbesondere das Präskriptionsargument. Denn einerseits werden Wissenschaftsanwenderinnen und -anwender typischerweise nicht in der Lage sein, die genauen Effekte eines Werturteils abzuschätzen. Noch unrealistischer scheint es, dass sie diese Werturteile durch ihre eigenen ersetzen und hieraus „alternative assessments and conclusions" ableiten können. Andererseits ermöglicht Transparenz den Anwenderinnen und Anwendern nur eine schwache Form von Autonomie („autonomy *qua* recipient", s. Teil III). Für eine ambitioniertere Form von Autonomie („autonomy *qua* author") ist dies jedoch nicht hinreichend. Obwohl Transparenz ein sinnvoller Ausgangspunkt für den Umgang mit Werturteilen ist, empfiehlt sich daher die Suche nach weiteren, über das bloße Explizit-Machen normativer Annahmen hinausgehenden Strategien.

4.2 STRATEGIEN II: DAS „WIE“ WERTBELADENER WISSENSCHAFT

Diese Strategien müssen zwei Aspekte klären: Zum einen das *Wie* von Werturteilen im Forschungsprozess, also die Situationen, Funktionen und Absichten legitimen Werturteilsgebrauchs; und zum anderen das *Was* der verwendeten Werturteile, also ihr substanzieller Gehalt sowie die ethischen, sozialen oder politischen Quellen, aus denen sie stammen[5]. Je nach Systematik lässt sich dabei eine große Zahl von Ansätzen identifizieren (s. auch Douglas, 2015; Holman & Wilholt, 2022). Angesichts der vielfältigen Begründungen und Kritiken des Wertfreiheitsideals ist dies insofern folgerichtig, als die verschiedenen Strategien jeweils andere Argumente für relevant erachten. Hinzu kommt, dass sich auch das Wertfreiheitsideal selbst als Antwort auf die Fragen nach dem Wie und dem Was von Werturteilen begreifen lässt. Ich werde diese für Vertreterinnen und Vertreter von Wertfreiheit typischen Strategien in Teil II vorstellen. Einige Ansätze der Wertfreiheitskritikerinnen und -kritiker werde ich außerdem in den Teilen III und IV diskutieren. Da ich mich dort jedoch auf bestimmte Strategien konzentriere, nämlich auf Ansätze aus dem Umfeld des Arguments des induktiven Risikos sowie auf den Ansatz Philip Kitchers, möchte ich zunächst einen etwas allgemeineren Überblick geben.

Eine vieldiskutierte Strategie zur Spezifizierung des Wie wertbeladener Wissenschaft stammt von Heather Douglas (2000, 2008, 2009). Danach sollen Werturteile den Grad der Sicherheit festlegen, der für eine wissenschaftliche Entscheidung angesichts ihrer außerwissenschaftlichen Folgen nötig ist. Douglas bezeichnet dies als „indirekte Rolle“. Demgegenüber seien Werturteile in einer „direkten Rolle“, nämlich als „reasons in themselves to accept or reject an empirical claim” (2008, S. 8), nicht akzeptabel. Eine etwas andere Position vertritt Daniel Steel (2010, 2017). Für Steel ist nicht die Rolle eines Werturteils ausschlaggebend, sondern sein Einfluss auf die wissenschaftliche Wahrheitssuche. Akzeptabel sind ethische und politische Werturteile aus seiner Sicht nur dann, wenn sie mit diesem Ziel vereinbar sind. Der bereits erwähnte aims approach erweitert dies um Ziele neben Wahrheit, so dass das Akzeptabilitätskriterium eines Werturteils in seiner Übereinstimmung mit den in einem gegebenen Forschungskontext relevanten epistemischen und nicht-epistemischen Zielen liegt (Brown, 2013; Elliott, 2013).

5 Wie ich in Teil III zeige, muss darüber hinaus das *relative Gewicht* der verschiedenen Werturteilsarten festgelegt werden. Da diese Frage jedoch bisher kaum diskutiert wird, gehe ich darauf hier nicht weiter ein.

Eine weitere Strategie betrachtet Werturteile als „Tiebreaker" für wissenschaftlich ebenbürtige Alternativen (eine ähnliche Position vertrete ich in Teil III). Hiernach besteht die legitime Funktion von Werturteilen darin, unter allen mit den verfügbaren Evidenzen kompatiblen Optionen diejenige auszuwählen, die aus ethischer Sicht am vorteilhaftesten erscheint (Longino, 1990; Kourany, 2003). Ein anderer Ansatz thematisiert die Absicht des Werturteilsgebrauchs. Nach diesem Ansatz sind Werturteile akzeptabel, solange sie nicht dazu missbraucht werden, erwünschte Ergebnisse zu erzielen und unerwünschte zu verhindern. So fordert etwa Elizabeth Anderson: „We need to ensure that value judgments do not operate to drive inquiry to a predetermined conclusion. This is our fundamental criterion for distinguishing legitimate from illegitimate uses of values in science" (2004, S. 11). Ein ähnlicher Ansatz bezieht sich ebenfalls auf die Absicht des Werturteilsgebrauchs, jedoch mit Blick auf die Rezipientinnen und Rezipienten eines Forschungsergebnisses. Hiernach sind Werturteile nur dann akzeptabel, wenn sie nicht verwendet werden, um das wissenschaftliche und außerwissenschaftliche Publikum über den wahren Charakter einer Forschungsentscheidung zu täuschen (Carrier, 2013).

4.3 STRATEGIEN III: DAS „WAS" WERTBELADENER WISSENSCHAFT

Wie steht es aber mit dem *Was* wertbeladener Wissenschaft, also der Frage, welche Werturteile konkret gefällt werden sollen? Eine mögliche Antwort hierauf lautet, dass Wissenschaftlerinnen und Wissenschaftler die *richtigen* Werte verwenden sollen, wobei die Richtigkeit eines Werturteils nach seiner Übereinstimmung mit wohlbegründeten ethischen Prinzipien bemessen wird. In der feministischen Wissenschaftsphilosophie werden hier etwa egalitäre (Kourany 2008), demokratische (Intemann, 2015) oder solche Werte vorgeschlagen, die der Emanzipation von Frauen und benachteiligten Gruppen dienen (Kourany, 2003; Longino, 2008). Ein anderer, ebenfalls aus dem Umfeld der feministischen Wissenschaftsphilosophie stammender Ansatz kommt ohne *substanzielle* ethische Prinzipien aus, schlägt aber eine Reihe *prozeduraler* Prinzipien vor. Nach diesem von Helen Longino (1990, 2002) vertretenen Ansatz können Wissenschaftlerinnen und Wissenschaftler je unterschiedliche Werturteile fällen, solange diese Gegenstand offener Diskurse innerhalb der wissenschaftlichen Gemeinschaft sind. Für diese Diskurse sollen wiederum bestimmte Regeln gelten, etwa das Prinzip der intellektuellen Gleichheit oder die Pflicht zur Aufnahme von Kritik (Longino, 2002, S. 128-135).

Eine weiterer Ansatz beruht auf der Tatsache, dass wissenschaftliche Disziplinen über implizite und explizite Methodenstandards verfügen (Wilholt, 2009, 2013). Werden diese Standards als Werturteile über die Akzeptabilität von Fehlerwahrscheinlichkeiten interpretiert, so lässt sich hieraus die Annahme ableiten, dass sich Wissenschaftlerinnen und Wissenschaftler *im Normalfall* an den kollektiven Werten ihrer Disziplin orientieren sollen. Abweichungen hiervon können nach diesem Ansatz zwar ebenfalls legitim sein, müssen aber explizit gemacht und begründet werden: „While there is nothing objectionable about researchers openly dissenting from the priorities inscribed in these conventions based on their own value judgments (for example, when they push for methodological reform), it is inappropriate to tacitly or secretly militate against them" (Holman & Wilholt, 2022, S. 213). Ein konzeptuell ähnlicher, aber mehr auf die Kommunikation mit außerwissenschaftlichen Gruppen gerichteter Vorschlag basiert auf der normativen Übereinstimmung mit einem Zielpublikum. Danach sollen sich Wissenschaftlerinnen und Wissenschaftler bei der Interpretation von Forschungsergebnissen an den Werturteilen orientieren, von denen sie plausibel erwarten können, dass sie in einer gegebenen Gruppe von Wissenschaftsrezipientinnen und -rezipienten geteilt werden (John, 2015, 2019).

Eine weitere, auch jenseits der Wissenschaftsphilosophie populäre Gruppe von Ansätzen beruht auf der Beteiligung außerwissenschaftlicher Stakeholder. Nach diesen Ansätzen sollen die verschiedenen Anspruchsgruppen wissenschaftlicher Forschung, etwa konkrete Nutzerinnen und Nutzer oder die allgemeine Öffentlichkeit, den Inhalt der fraglichen Werturteile selbst festlegen. In der Wissenschaftsphilosophie wird dies etwa von Philip Kitcher (2011) vertreten (ich diskutiere Kitchers Vorschlag in Teil IV). Darüber hinaus existiert ein breites Spektrum ähnlicher Ansätze aus anderen Kontexten, etwa den Science and Technology Studies (z.B. Gibbons et al., 1994), der Nachhaltigkeitsforschung (z.B. Mauser et al., 2013) oder der klimawissenschaftlichen Politikberatung (Edenhofer & Kowarsch, 2015). Das Spektrum denkbarer Beteiligungsformate ist dabei umfangreich und umfasst etwa Bürgerdeliberationen, Workshops, Interviews oder Befragungen. Eng damit verwandt ist der ebenfalls in den Klimawissenschaften populäre Szenarioansatz. Danach sollen normative Annahmen systematisch variiert werden (Frisch, 2013), um so ein Set konditionaler, möglichst viele Werthaltungen abdeckender Szenarien zu erhalten (Held, 2011). Obwohl dieser Ansatz weniger partizipativ und stärker auf bestimmte Annahmen begrenzt ist (etwa der Grad an Optimismus hinsichtlich zukünftiger Entwicklungen), orientiert sich der Inhalt von Werturteilen auch hier an außerwissenschaftlichen Stakeholdern.

5. Überblick über diese Arbeit

Die vorangegangenen Kapitel haben die Vielfalt der Argumente und Ansätze illustriert, die im Zusammenhang mit dem Wertfreiheitsthema diskutiert werden. Gleichzeitig ist klar geworden, dass das Thema keineswegs als „playground for intellectuals" (Kowarsch, 2016, S. 6) verstanden werden kann. Vielmehr stehen wichtige epistemische Güter wie Wahrheit oder logische Schlüssigkeit, aber auch nicht-epistemische Güter wie die Autonomie außerwissenschaftlicher Stakeholder oder die Legitimität wissenschaftlicher Politik- und Gesellschaftsberatung auf dem Spiel. Die scheinbar theoretische Frage nach der Rolle von Werturteilen im Forschungsprozess hat somit praktische Folgen für unser Verständnis von *guter Wissenschaft* und *guten wissenschaftsgestützten Entscheidungen*.

Die vorliegende Arbeit beinhaltet drei Forschungsartikel und einen konzeptionellen Aufsatz, in denen ich diese Frage und ihre Konsequenzen aus verschiedenen Blickwinkeln betrachte. Die einzelnen Teile sind thematisch fokussiert, d.h. sie betrachten nicht alle der skizzierten Argumente und Ansätze, sondern konzentrieren sich auf bestimmte Aspekte des Wertfreiheitsthemas. Alle vier Teile sind dabei als eigenständige Debattenbeiträge zu verstehen und können separat gelesen werden. Bei den Teilen III, IV und V handelt es sich um englischsprachige Beiträge für Fachzeitschriften. Eine Ausnahme bildet Teil II. Hierbei handelt es sich um einen deutschsprachigen Aufsatz, der mehr auf Begriffsklärung ausgerichtet ist und einen stärkeren Überblickscharakter hat als die restlichen Teile. Auch dieser Teil kann jedoch unabhängig von der übrigen Arbeit gelesen werden. Aufgrund dieses Aufbaus sind die Einzelteile stärker zugespitzt als es in einer monografischen Abhandlung möglich wäre; gleichzeitig bringt der kumulative Aufbau kleinere Inkonsistenzen und Redundanzen mit sich. So beinhalten alle vier Teile jeweils eigene Rekonstruktionen des Wertfreiheitsideals und der relevanten Argumente. Die unterschiedlichen Kontexte bedingen dabei leicht abweichende Terminologien und Darstellungsweisen. Um dennoch eine gewisse Kohärenz zu gewährleisten, sind den Teilen anstatt der originalen Abstracts kurze deutschsprachige Zusammenfassungen vorangestellt, die als thematische Brücken zwischen den jeweiligen Diskussionskontexten fungieren.

Der *zweite Teil* dieser Arbeit besteht in dem Aufsatz „Das Wertfreiheitsideal: Bedeutung, Grenzen und Kritik eines komplexen Begriffs". Ziel des Aufsatzes ist es, ein belastbares Verständnis des Wertfreiheitsideals zu erarbeiten und dieses in der klassischen und zeitgenössischen Literatur zu verorten. Wie ich zeigen werde, ist das Wertfreiheitsideal

eine anspruchsvolle, aus mehreren Teilthesen und Einschränkungen bestehende philosophische Position. Die einzelnen Begriffselemente ermöglichen ihrerseits je unterschiedliche Lesarten, von denen wiederrum abhängt, welche Arten von Kritik an das Wertfreiheitsideal herangetragen werden können. Der Aufsatz stellt diese Lesarten vor und prüft sie auf ihre Plausibilität. Ein Schwerpunkt liegt dabei auf dem Sollens- und dem Möglichkeitsanspruch des Ideals. So kläre ich etwa, ob Wertfreiheit als *bloßes Ideal* verstanden werden kann, ob dieses Ideal *im Prinzip* erreichbar sein muss, inwiefern das Ideal *Verpflichtungscharakter* hat und ob es selbstwidersprüchlich ist, Wertfreiheit einen *Wert* beizumessen.

Im *dritten Teil* dieser Arbeit, dem Forschungsartikel „Inductive Risk: Does it Really Refute Value-Freedom?", prüfe ich eines der wichtigsten Argumente gegen das Wertfreiheitsideal: das Argument des induktiven Risikos. Dabei diskutiere ich die bereits erwähnte deskriptive und normative Lesart dieses Arguments; außerdem betrachte ich das ebenfalls erwähnte Präskriptionsargument sowie eine (terminologisch und formal leicht abgewandelte) Version des Sein-Sollen-Arguments. Der technische Kern des Teils besteht in einem idealisierten Bayesianischen Entscheidungssetting. Auf dieser Grundlage kläre ich, ob und in welchem Sinn das Argument des induktiven Risikos eine *prinzipielle* Widerlegung des Wertfreiheitsideals ermöglicht. Ich zeige, dass dies tatsächlich der Fall ist, jedoch nur in normativer Hinsicht und begrenzt auf *Tiebreaker*-Situationen. Allerdings müssen diese Situationen anders als in der Literatur üblich interpretiert werden, nämlich als Entscheidungen zwischen Optionen mit gleichem oder ähnlichem Erwartungsnutzen. Weiterhin zeige ich, wie derartige Entscheidungen wertbeladen, aber dennoch nicht-präskriptiv und logisch unproblematisch sein können. Am Ende des Teils argumentiere ich, dass die Bayesianische Rekonstruktion trotz ihres idealisierenden Charakters informativ für tatsächliches Forschungshandeln ist.

Der *vierte Teil* beinhaltet den Forschungsartikel „Citizen Participation in Kitcher's Well-Ordered Science: What Kind of Ideal Is Deliberation?". Hier diskutiere ich den erwähnten Ansatz Philip Kitchers. Dieser beruht auf einem *deliberativen Ideal*, d.h. auf einer Reihe prozeduraler Regeln, die die Akzeptabilität wertbeladener Forschungsentscheidungen an die (tatsächliche oder hypothetische) Durchführung einer inklusiven, egalitären und rationalen Diskussion unter außerwissenschaftlichen Stakeholdern bindet. Wie Kitcher jedoch selbst betont, ist dieses Ideal nicht umsetzbar. Sein Vorschlag setzt sich somit der *realistischen Kritik* aus, dass unerreichbare Ziele schlechte Ziele sind. In der Tat entspringt diese Kritik aus Kitchers eigener Zurückweisung des Wertfreiheitsideals. Es stellt sich daher die Frage, ob das deliberative Ideal denselben realistischen

Argumenten zum Opfer fällt, wie sie Kitcher gegen das Ideal der Wertfreiheit vorträgt. Um diese Frage zu beantworten, führe ich vier Interpretationen des Begriffs „Ideal" ein. Ich argumentiere, dass die aussichtsreichste Verteidigung des deliberativen Ideals in dem besteht, was ich die *rekonstruktive Lesart* nenne.

Der *fünfte Teil*, der Forschungsartikel „Models of Science and Society: Transcending the Antagonism", nimmt eine Sonderstellung ein. Er behandelt das Problem, wie konkurrierende Antworten auf die Frage nach dem angemessenen Wissenschafts-Gesellschafts-Verhältnis systematisiert und produktiv aufeinander bezogen werden können. Anders als die restliche Arbeit ist dieser Teil interdisziplinär und anwendungsorientiert. Weiterhin bezieht er sich auf eine Diskussion, die eher in den Science and Technology Studies und der transdisziplinären Forschung geführt wird. Das Wertfreiheitsthema ist jedoch auch hier zentral. Eine Parallele mit der philosophischen Debatte besteht außerdem in der populären Ablehnung traditioneller Wissenschaftsverständnisse. Im Gegensatz zu dieser mitunter polemisch vorgetragenen Kritik argumentiere ich, dass konkurrierende Modelle des Wissenschafts-Gesellschafts-Verhältnisses koexistieren können, wenn sie angemessen interpretiert werden. Hierfür muss das verbreitete Verständnis, nach dem diese Modelle widerstreitende Lager oder Akteursüberzeugungen repräsentieren, durch ein idealtypisches und heuristisches Modellverständnis ersetzt werden. An einem konkreten Beispiel zeige ich, wie dieses Verständnis in realen Wissenschafts-Gesellschafts-Interaktionen genutzt werden kann. Trotz der anderen Ausrichtung spiegelt dieser Teil damit die Motivation der Gesamtarbeit wider – einer Motivation, die Robert Proctor vor drei Jahrzehnten zum Ausdruck gebracht hat und die heute aktueller denn je ist:

> It has become fashionable to claim that "all science is social," and there are many good reasons for this fashion. Still, we should not forget that there are some good reasons that scholars have tried (and continue to try) to distance their work from questions of values or politics. […] *My hope is that the problem of whether science is or ought to be value-free will remain an open question*, a question of how properly to politicize or depoliticize science, to enliven science education, to bring science more closely and humanely to bear on matters of vital human interest (Proctor, 1991, S. x-xi, meine Hervorh.).

Literatur

Anderson, E. (2004). Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce. *Hypathia*, 19(1), 1–24. https://doi.org/10.1111/ j.1527-2001.2004.tb01266.x

Arendt, H. (1994). Über die Revolution. Piper. (Erstveröffentlichung 1963)

Bacon, F. (2017). *Neues Organon. Große Erneuerung der Wissenschaften.* Contumax. (Erstveröffentlichung 1620)

Bělohrad, R. (2011). The Is-Ought Problem, the Open Question Argument, and the new science of morality. *Human Affairs*, 21, 262. https://doi.org/10.2478/s13374-011-0027-3

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, 3(2), 207–220. https://doi.org/10.1007/s13194-012-0062-x

Biddle, J., & Kukla, R. (2017). The Geography of Epistemic Risk. In K. Elliott, & T. Richards (Hrsg.), *Exploring Inductive Risk: Case Studies of Values in Science* (S. 215–237). Oxford University Press.

Biddle, J., & Winsberg, E. (2010). Value Judgements and the Estimation of Uncertainty in Climate Modelling. In P.D. Magnus, & J. Busch (Hrsg.), *New Waves in Philosophy of Science* (S. 172–197). Palgrave Macmillan.

Bright, L. (2018). Du Bois' democratic defence of the value free ideal. *Synthese*, 95(5), 2227–2245. https://doi.org/10.1007/s11229-017-1333-z

Brown, M. J. (2013). Values in Science beyond Underdetermination and Inductive Risk. *Philosophy of Science*, 80, 829–839. https://doi.org/10.1086/673720

Campbell, T. (1970). The Normative Fallacy. *The Philosophical Quarterly*, 20(81), 368–377. https:// doi.org/10.2307/2217655

Carrier, M. (2013). Values and objectivity in science: Value-ladenness, pluralism and the epistemic attitude. *Science & Education*, 22, 2547–2568. https://doi.org/10.1007/s11191-012-9481-5

Collins, H. (1985). *Changing order. Replication and induction in scientific practice*. Sage.

Dewey, J. (1939). *Theory of Valuation*. University of Chicago Press.

Douglas, H. E. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, 67(4), 559–579. https://doi.org/10.1086/392855

Douglas, H. (2008). The Role of Values in Expert Reasoning. *Public Affairs Quarterly*, 22(1), 1–18.

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Douglas, H. E. (2015), Values in Science. In P. Humphreys (Hrsg.), *The Oxford Handbook of Philosophy of Science* (S. 609–630). Oxford University Press.

D'Arms, J. & Jacobson, D. (2000). The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophy and Phenomenological Research*, 61(1), 65–90. https://doi.org/10.2307/2653403

Duhem, P. (1954). *The aim and structure of physical theory*. Princeton University Press. (Erstveröffentlichung 1906)

Dupré, J. (2013). Tatsachen und Werte. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 255–271). Suhrkamp.

Edenhofer, O., & Kowarsch, M. (2015). Cartography of pathways: A new model for environmental policy assessments. *Environmental Science & Policy*, 51, 56–64. https://doi.org/10.1016/j.envsci. 2015.03.017

Elliott, K. C. (2013). Douglas on values: From indirect roles to multiple goals. *Studies in History and Philosophy of Science*, 44, 375–383. https://doi.org/10.1016/j.shpsa.2013.06.003

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic Values and the Multiple Goals of Science. *Philosophy of Science*, 81(1), 1–21. https://doi.org/10.1086/674345

Elliott, K. C., & Resnik, D. B. (2014). Science, policy, and the transparency of values. *Environmental Health Perspectives*, 122(7), 647–650. https://doi.org/10.1289/ehp.1408107

Elliott, K. C., McCright, A. M., Allen, S., & Dietz, T. (2017). Values in environmental research: Citizens' views of scientists who acknowledge values. *PloS One*, 12(10), e0186049. https://doi.org/ 10.1371/journal.pone.0186049

Elliott, K. C., & Steel, D. (2017). Introduction: Values and Science: Current Controversies. In K. Elliott, & D. Steel (Hrsg.), *Current Controversies in Values and Science* (S. 1–11). Routledge.

Fleck, L. (1980): *Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv*. Suhrkamp. (Erstveröffentlichung 1935)

Frisch, M. (2013). Modeling Climate Policies: A Critical Look at Integrated Assessment Models. *Philosophy & Technology*, 26, 117–137. https://doi.org/10.1007/s13347-013-0099-6

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge. The Dynamics of Science and Research in Contemporary Societies*. Sage.

Giere, R. N. (2003): A New Program for Philosophy of Science? *Philosophy of Science*, 70(1), 15-21. https://doi.org/10.1086/367865

Guevara, D. (2008). Rebutting formally valid counterexamples to the Humean "is-ought" dictum. *Synthese*, 164, 45–60. https://doi.org/10.1007/s11229-007-9215-4

Haack, S. (2003). Knowledge and Propaganda. Reflections of an Old Feminist. In C. Pinnick, N. Koertge, & R. Almeder (Hrsg.), *Scrutinizing Feminist Epistemology. An Examination of Gender in Science* (S. 7–19). Rutgers University Press.

Habermas, J. (2013): Erkenntnis und Interesse. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 57–73). Suhrkamp. (Erstveröffentlichung 1965)

Harding, S. (1995). "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3), 331–349. https://doi.org/10.1007/BF01064504

Hedgecoe, A. M. (2004). Critical bioethics: beyond the social science critique of applied ethics. *Bioethics*, 18(2), 120–143. https://doi.org/10.1111/j.1467-8519.2004.00385.x

Held, H. (2011). Dealing with Uncertainty – From Climate Research to Integrated Assessment of Policy Options. In G. Gramelsberger, & J. Feichter (Hrsg.), *Climate Change and Policy. The Calculability of Climate Change and the Challenge of Uncertainty* (S. 113–126). Springer.

Hempel, C. (1965). Science and Human Values. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science* (S. 81–96). The Free Press.

Hicks, D. J. (2014). A new direction for science and values. *Synthese*, 191, 3271–3295. https:// doi.org/10.1007/s11229-014-0447-9

Holman, B., & Wilholt, T. (2022). The New Demarcation Problem. *Studies in History and Philosophy of Science*, 91, 211–220. https://doi.org/10.1016/j.shpsa.2021.11.011

Hoyningen-Huene, P. (2006). Context of Discovery Versus Context of Justification and Thomas Kuhn. In J. Schickore, & F. Steinle (Hrsg.), *Revisiting Discovery and Justification. Historical and philosophical perspectives on the context distinction* (S. 119–131). Springer.

Hume, D. (1978): *Ein Traktat über die menschliche Natur*. Bd. 2. Meiner (Erstveröffentlichung 1740)

Intemann, K. (2005). Feminism, underdetermination, and values in science. *Philosophy of science*, 72(5), 1001–1012. https://doi.org/10.1086/508956

Intemann, K. (2015). Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science*, 5, 217–232. https://doi.org/10.1007/s13194-014-0105-6

Jeffrey, R. C. (1956). Valuation and acceptance of scientific hypotheses. *Philosophy of Science*, 22, 237–246. https://doi.org/10.1086/287489

John, S. (2015). Inductive risk and the contexts of communication. *Synthese*, 192(1), 79–96. https://doi.org/10.1007/s11229-014-0554-7

John, S. (2019). Science, truth and dictatorship: Wishful thinking or wishful speaking?. *Studies in History and Philosophy of Science*, 78, 64–72. https://doi.org/10.1016/j.shpsa.2018.12.003

Kitcher, P. (2011). *Science in a democratic society*. Prometheus Books.

King, D., Schrag, D., Zhou, D., Ye, Q., & Ghosh, A. (2015). *Climate Change: A Risk Assessment*. University of Cambridge, Centre for Science and Policy. https://www.csap.cam.ac.uk/media/uploads/files/1/climate-change--a-risk-assessment-v11.pdf

Koertge, N. (2013). Wissenschaft, Werte und die Werte der Wissenschaft. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 233–251). Suhrkamp.

Kourany, J. A. (2003). A Philosophy of Science for the Twenty-First Century. *Philosophy of Science*, 70(1), 1–14. https://doi.org/10.1086/367864

Kourany, J. A. (2008). Replacing the Ideal of Value-Free Science. In M. Carrier, D. Howard, & J. Kourany (Hrsg), *The Challange of the Social and the Pressure of Practice. Science and Values Revisited* (S. 87–109). University of Pittsburgh Press.

Kowarsch, M. (2016). *A Pragmatist Orientation for the Social Sciences in Climate Policy. How to Make Integrated Economic Assessments Serve Society.* Springer.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, Value Judgement, and Theory Choice. In *The Essential Tension. Selected Studies in Scientific Tradition and Change* (S. 320–339). University of Chicago Press.

Lacey, H. (1999). *Is Science Value Free? Values and Scientific Understanding*. Routledge.

Latour, B., & Woolgar, S. (1979). *Laboratory life. The social construction of scientific facts*. Sage Publications.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Longino, H. E. (2004): How Values Can Be Good for Science. In P. Machamer, & G. Wolters (Hrsg.), *Science, Values, and Objectivity* (S. 127–142). University of Pittsburgh Press.

Longino, H. E. (2008). Values, Heuristics, and the Politics of Knowledge. In M. Carrier, D. Howard, & J. Kourany (Hrsg.), *The Challenge of the Social and the Pressure of the Practice* (S. 68–86). University of Pittsburgh Press.

Mauser, W., Klepper, G., Rice, M., Schmalzbauer, B. S., Hackmann, H., Leemans, R., & Moore, H. (2013): Transdisciplinary global change research: the co-creation of knowledge for sustainability. *Current Opinion in Environmental Sustainability*, 5, 420–431.

Luhmann, N. (1990). *Die Wissenschaft der Gesellschaft*. Suhrkamp.

Merton, R. K. (1973). The normative structure of science. In *The sociology of science. Theoretical and empirical investigations* (S. 267–278). University of Chicago Press. (Erstveröffentlichung 1942)

MacIntyre, A. (2013): *After Virtue: A Study in Moral Theory*. Bloomsbury. (Erstveröffentlichung 1981)

McMullin, E. (1982). Values In Science. *Proceedings of the Biennial Meeting of the Philosophy of Science Association. Vol. Two: Symposia and Invited Papers*, 3–28. https://doi.org/10.1086/psaprocbienmeetp.1982.2.192409

Pidgen, C. (2010). On the Triviality of Hume's Law: a Reply to Gerhard Schurz. In C. Pidgen (Hrsg.), *Hume on Is and Ought* (217–238). Palgrave Macmillan.

Pidgen, C. (2016). Hume on Is and Ought. In P. Russel (Hrsg.), *The Oxford Handbook of Hume* (S. 401–415). Oxford University Press.

Polanyi, M. (1962). The Republic of science. *Minerva*, 1(1), 54–73. https://doi.org/10.1007/ BF01101453

Popper, K. (1974). Die Logik der Sozialwissenschaften. In T. Adorno, H. Albert, R. Dahrendorf, J. Habermas, H. Pilot, & K. Popper (Hrsg.), *Der Positivismusstreit in der deutschen Soziologie* (S. 103–123). Luchterhand.

Prior, A. (1960). The autonomy of ethics. *Australasian Journal of Philosophy*, 38(3), 199–206.

Proctor, R. (1991). *Value-free science? Purity and power in modern knowledge*. Harvard University Press.

Putnam, H. (2002). *The Collapse of the Fact/value Dichotomy and Other Essays*. Harvard University Press.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, 60(1), 20–43.

Reichenbach, H. (1961). *Experience and Prediction*. University of Chicago Press. (Erstveröffentlichung 1938)

Rittel, H. W. J., & Webber, M. M. (1973): Dilemmas in a general theory of planning. *Policy Science*, 4, 155–169. https://doi.org/10.1007/BF01405730

Rudner, R. (1953). The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science*, 20(1), 1–6. https://doi.org/10.1086/287231

Schneider, S.H. (1997). Integrated assessment modeling of global climate change: Transparent rational tool for policy making or opaque screen hiding value-laden assumptions?. *Environmental Modeling & Assessment,* 2, 229–249. https://doi.org/10.1023/A:1019090117643

Schurz, G. (1997). *The Is-Ought-Problem. An Investigation in Philosophical Logic.* Springer.

Schurz, G. (2011). *Evolution in Natur und Kultur: Eine Einführung in die verallgemeinerte Evolutionstheorie*. Spektrum.

Schurz, G. (2013). Wertneutralität und hypothetische Werturteile in den Wissenschaften. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 305–334). Suhrkamp.

Searle, J. (1964). How to Derive "Ought" From "Is". *The Philosophical Review*, 73(1).

Singer, D. (2015): Mind the Is Ought Gap. *Journal of Philosophy,* 112(4), 193–210. https://doi.org/ 10.5840/jphil2015112412

Stanton, E.A. (2011). Negishi welfare weights in integrated assessment models: the mathematics of global inequality. *Climatic Change,* 107, 417–432. https://doi.org/10.1007/s10584-010-9967-6

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, 77, 14–34. https://doi.org/10.1086/650206

Steel, D. (2016). Climate Change and Second-Order Uncertainty: Defending a Generalized, Normative, and Structural Argument from Inductive Risk. *Perspectives on Science*, 24(6), 696–721.

Steel, D. (2017). Qualified Epistemic Priority. Comparing Two Approaches to Values in Science. In K. Elliott, & D. Steel (Hrsg.), *Current Controversies in Values and Science* (S. 49–63). Routledge.

Strohschneider, P. (2014). Zur Politik der Transformativen Wissenschaft. In A. Brodocz, D. Herrmann, R. Schmidt, D. Schulz, & J. Schulze Wessel (Hrsg.), *Die Verfassung des Politischen* (S. 175– 192). Springer.

Weber, M. (1988). Die „Objektivität" sozialwissenschaftlicher und sozialpolitischer Erkenntnis. In *Gesammelte Aufsätze zur Wissenschaftslehre* (7. Aufl., hrsg. v. Johannes Winckelmann) (S. 146–214). Mohr. (Erstveröffentlichung 1904)

Weber, M. (1988). Der Sinn der „Wertfreiheit“ der soziologischen und ökonomischen Wissenschaften. In *Gesammelte Aufsätze zur Wissenschaftslehre* (7. Aufl., hrsg. v. J. Winckelmann) (S. 489–540). Mohr. (Erstveröffentlichung 1917)

Weber, M. (1988). Wissenschaft als Beruf. In *Gesammelte Aufsätze zur Wissenschaftslehre* (7. Aufl., hrsg. v. Johannes Winckelmann) (S. 582–613). Mohr. (Erstveröffentlichung 1919)

Weingart, P. (2001). *Die Stunde der Wahrheit? Zum Verhältnis der Wissenschaft zu Politik, Wirtschaft und Medien in der Wissensgesellschaft*. Velbrück.

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science*, 40(1), 92–101. https://doi.org/10.1016/j.shpsa.2008.12.005

Wilholt, T. (2013). Epistemic Trust in Science. *British Journal for the Philosophy of Science*, 64, 233–253. https://doi.org//10.1093/bjps/axs007

Williams, B.A.O. (1985). *Ethics and the limits of philosophy*. Harvard University Press.

Winsberg, E. (2012). Values and Uncertainties in the Prediction of Global Climate Models. *Kennedy Institute of Ethics Journal*, 22(2), 111–137. https://doi.org/10.1353/ken.2012.0008

Winsberg, E., Huebner, B., & Kukla, R. (2014). Accountability and values in radically collaborative research. S*tudies in History and Philosophy of Science* 46 (2014) 16–23. http://dx.doi.org/10.1016/ j.shpsa.2013.11.007

II. DAS WERTFREIHEITSIDEAL: BEDEUTUNG, GRENZEN UND KRITIK EINES KOMPLEXEN BEGRIFFS

ZUSAMMENFASSUNG Im ersten Teil habe ich einen Überblick über die philosophische Wertfreiheitsdebatte gegeben. Dabei sind jedoch wichtige Aspekte offen geblieben, etwa die Bedeutung der normativen und deskriptiven Geltungsansprüche des Wertfreiheitsideals oder die Wissenschaftsbereiche und Werturteilstypen, auf die es sich bezieht. Im zweiten Teil kläre ich diese Fragen anhand einer systematischen Begriffsanalyse. Gleichzeitig verorte ich die einzelnen Begriffselemente in der klassischen und zeitgenössischen Literatur. Ich rekonstruiere das Wertfreiheitsideal als Menge von vier Teilthesen, die ihrerseits durch mehrere Spezifikationen und Einschränkungen gekennzeichnet sind. Dabei stelle ich mehrere Lesarten vor und führe einige neue, in der Wertfreiheitsdiskussion bisher noch nicht verwendete Unterscheidungen ein. Ich zeige, dass das Wertfreiheitsideal auf einer Strategie der *Einschränkung durch Abgrenzung* beruht – wobei es, wie ich ebenfalls zeige, einige wichtige Ausnahmen von dieser Strategie gibt. Dies wirft ein Licht auf die im ersten Teil skizzierten Kritiken des Wertfreiheitsideals und verdeutlicht, welche Optionen Gegnerinnen und Gegnern des Wertfreiheitsideals prinzipiell zur Verfügung stehen.

1. Einleitung

Was ist wissenschaftliche Wertfreiheit – und was bedeutet es, Wertfreiheit als „Ideal" zu vertreten? Angesichts der langen und kontroversen Geschichte der Wertfreiheitsdebatte (Proctor, 1991; Dahms, 2013) ist es nicht verwunderlich, dass Begriffe wie „Werturteil", „Wertfreiheit" und „Wertfreiheitsideal" keineswegs immer klar sind. Ein Grund hierfür ist die Vielzahl von Aspekten, die im Rahmen dieser Debatte diskutiert werden. So konzentriert sich ein Teil der Debatte auf Wertfreiheit als *Ideal* (z.B. Kourany, 2008; Koertge, 2013; Bright, 2018), während ein anderer Teil die *Möglichkeit* von Wertfreiheit thematisiert (z.B. Rudner 1953; Lacey, 1999; Biddle & Winsberg, 2010); ein weiterer Schwerpunkt liegt auf den *Bereichen* oder Kontexten von Wissenschaft, für die das Wertfreiheitsideal Geltung beansprucht (Reichenbach 1938/1961; Hoyningen-Huene, 2006; Bueter, 2015); und nicht zuletzt stellt sich die Frage, wovon überhaupt die Rede ist, wenn in dieser Debatte von *Werten* gesprochen wird (z.B. Kuhn, 1977; Longino, 1996; Machamer & Douglas, 1999). Hierdurch entsteht eine begriffliche Komplexität, die das Wertfreiheitsideal – und damit die Thesen, auf die seine Vertreterinnen und Vertreter festgelegt sind – zu einer erklärungsbedürftigen Position machen. Insbesondere muss geklärt werden, welche Implikationen sich aus der Zurückweisung oder Verteidigung einer der Teilthesen des Wertfreiheitsideals für das Ideal im Ganzen ergeben.

In dem vorliegenden Aufsatz greife ich diese Komplexität auf und versuche, ein umfassendes Verständnis des Wertfreiheitsideals, seiner Teilaspekte und der zugrunde liegenden Begriffe zu erarbeiten. Ich rekonstruiere das Wertfreiheitsideal als Menge von vier Teilthesen, die ihrerseits eine Reihe begrifflicher Spezifikationen enthalten. Hierfür führe ich mehrere Unterscheidungen sowie unterschiedliche Lesarten der Teilthesen ein. Ziel ist es, das Wertfreiheitsideal in seiner Vielschichtigkeit zu begreifen und die verschiedenen Begriffsverständnisse voneinander abzugrenzen. Gleichzeitig verorte ich diese Begriffsverständnisse in der klassischen und zeitgenössischen Literatur, wobei ich mich (abgesehen von wenigen Ausnahmen) auf die wissenschaftsphilosophische Diskussion beschränke. Thema der folgenden Abhandlung ist allerdings nicht, *ob* Wertfreiheit ein gutes Ideal ist und wodurch es, falls es nicht aufrechterhalten werden kann, *ersetzt* werden könnte. Diese Fragen diskutiere ich an anderer Stelle (Teil III und Teil IV dieser Arbeit). Im vorliegenden Zusammenhang geht es mir zunächst um die Begriffsarbeit.

Im ersten Kapitel führe ich, ausgehend von einer Unterscheidung in *input*- und *output-zentrierte* Fragestellungen, die zentralen Begriffe der Wertfreiheitsdebatte und die vier

Teilthesen des Wertfreiheitsideals ein. Ich unterscheide eine *normative*, eine *deskriptive*, eine *Kontextualitäts-* und eine *Differenzialitätsthese.* Weiterhin diskutiere ich das Verhältnis der Teilthesen untereinander sowie die Frage, ob Wertfreiheit als *bloßes Ideal* verstanden werden kann. Im zweiten Kapitel diskutiere ich zunächst den Verpflichtungscharakter des Wertfreiheitsideals, um dann die Frage zu klären, ob das Wertfreiheitsideal aufgrund seines normativen Charakters selbstwidersprüchlich ist. Das dritte Kapitel fragt, was es bedeutet, dass Wertfreiheit *im Prinzip* möglich sein soll und in welchem Umfang das Wertfreiheitsideal diesen Anspruch aufstellt. Das fünfte Kapitel thematisiert die dem Wertfreiheitsideal zugrunde liegenden Unterscheidungen zwischen wertfreien und wertbeladenen Wissenschaftsbereichen sowie zwischen epistemischen und nicht-epistemischen Werturteilen.

2. Was ist Wertfreiheit?

2.1 HISTORISCHER HINTERGRUND: INPUT- UND OUTPUTZENTRIERTE PROBLEMSTELLUNGEN

Die moderne Wertfreiheitsdebatte hat ihren Ursprung im frühen 20. Jahrhundert und konzentrierte sich zunächst auf die Frage, ob die damals jungen Disziplinen der Soziologie und Nationalökonomie praktische Handlungsempfehlungen aussprechen sollten (Dahms, 2013; Schurz & Carrier, 2013)[1]. Im Rahmen dieser ursprünglichen Problemstellung argumentierte etwa Max Weber, bis heute einer der prominentesten Befürworter wertfreier Wissenschaft, „daß es niemals Aufgabe einer Erfahrungswissenschaft sein kann, bindende Normen und Ideale zu ermitteln, um daraus für die Praxis Rezepte ableiten zu können“ (1904/1988, S. 149). Eine empirische Wissenschaft, so Weber, „vermag niemanden zu lehren, was er *soll*, sondern nur, was er *kann* und – unter Umständen – was er *will*“ (ebd., S. 151, Hervorh. i. Orig.). Folglich sei es falsch, dass „die Nationalökonomie Werturteile […] produziere und zu produzieren habe“ (ebd., S. 149). Wie sich an diesen Zitaten zeigt, bezog sich die ursprüngliche, manchmal als *erster Werturteilsstreit* bezeichnete Diskussion vorrangig auf das, was heute als wissenschaftliche Politikberatung (s. z.B. Maasen & Weingart, 2005; Falk et al., 2006) oder als Gesellschaftsberatung (Leggewie, 2006) bezeichnet wird. Die zentrale Fragestellung bestand darin, ob Wissenschaft Werturteile ethischer, sozialer oder politischer Art generieren kann und soll,

1 Die konzeptionellen Wurzeln des Wertfreiheitsbegriffs reichen jedoch mindestens bis in die frühe Neuzeit (Büter, 2012) oder gar bis in die antike Philosophie zurück (Proctor, 1991).

und in welcher Form diese dann in gesellschaftliche Sphären wie die Politik eingebracht werden sollen. Da sich diese Diskussion hauptsächlich auf die Resultate oder Implikationen wissenschaftlicher Forschung bezieht, kann man von einer *outputzentrierten* Fragestellung sprechen[2].

Aktuelle Debatten über Werturteile in wissenschaftlicher Politikberatung, etwa in der Klimapolitik (z.B. Hulme, 2009; Pielke, 2012; Edenhofer & Kowarsch, 2015) oder im Umgang mit COVID-19 (z.B. Martin et al. 2020), zeigen die ungebrochene Aktualität der outputzentrierten Frage. Dennoch hat sich die Diskussion in der Wissenschaftsphilosophie, aber auch in der sozialwissenschaftlichen und historischen Wissenschaftsforschung, ab den 50er und 60er Jahren des letzten Jahrhunderts zunehmend *inputzentrierten* Themen zugewandt (s. etwa Rudner, 1953; Kuhn, 1962; Hempel, 1965; Feyerabend, 1975; Popper, 1974; Latour & Woolgar, 1979; Collins, 1985; Longino 1990). Im Zentrum steht hier die Frage, welche Rolle Werturteile im Forschungsprozess selbst – also bevor Forschungsergebnisse produziert und in gesellschaftliche Diskurse eingebracht werden – spielen und spielen sollen. Thema dieser Diskussion ist somit weniger die *Generierung* von Werturteilen *durch* Wissenschaft, sondern um die *Verwendung* von Werturteilen *in* der Wissenschaft. Insbesondere in der Wissenschaftsphilosophie setzt sich dieser Themenfokus bis in die jüngere Debatte fort (z.B. Lacey, 1999; Douglas, 2000; Longino, 2002; Wilholt, 2009; Betz, 2013; Biddle, 2013).

Die Input/Output-Unterscheidung ist hilfreich, um diejenigen Werturteile, die womöglich in den Forschungsprozess eingebracht werden, von denjenigen zu unterscheiden, die womöglich aus ihm entspringen. Allerdings ist die Unterscheidung analytischer Natur und sollte nicht als substanzielle Aussage über die Natur dieser Werturteile missverstanden werden (d.h. die Unterscheidung spezifiziert Arten von Fragen, ohne die Antworten vorwegzunehmen; eine ähnliche Sichtweise hat Paul Hoyningen-Huene zur Unterscheidung zwischen Entdeckungs- und Rechtfertigungskontexten vorgeschlagen, s. 2006, S. 128-130). Auch sollte der Inputfokus der aktuellen Debatte eher als Schwerpunktsetzung anstatt als vollständige Abkehr von outputzentrierten Themen gedeutet werden[3]. Diese Schwerpunktsetzung hat jedoch wichtige Konsequenzen: Wenn

2 Die Schwerpunktsetzung auf den normativen Output von Wissenschaft bedeutet jedoch nicht, dass Fragen des normativen Inputs im ersten Werturteilsstreit gar keine Rolle gespielt haben. So widmete sich etwa Weber (1904/1988) ausführlich dem Fällen von Werturteilen zu Beginn und während des Forschungsprozesses (ebd., S. 170–171; S. 184; S. 192–193; S. 213–214). Dennoch lag der Debattenschwerpunkt auf der Frage, ob Werturteile das „*Produkt* fortschreitenden Erfahrungswissens sein können" (ebd., S. 154, Hervorh. i. Orig.).

3 So hat etwa Stephen John mehrere Beiträge zur Verwendung von Werturteilen in öffentlicher und politischer

in wissenschaftsphilosophischen Kontexten von Wertfreiheit oder -beladenheit die Rede ist, sind meist Aspekte des Forschungsprozesses (Datensammlung, Modellwahl, Hypothesenprüfung etc.) und nicht in erster Linie Fragen der Politik- und Gesellschaftsberatung gemeint.

2.2 WERTFREIHEIT: BEGRIFFE UND TEILTHESEN

Wie lässt sich nun der Begriff der Wertfreiheit in diesem inputzentrierten Sinne verstehen? Um die wissenschaftsphilosophische Wertfreiheitsdebatte und insbesondere die dort diskutierten Kritiken des Wertfreiheitsideals besser einordnen zu können, schlage ich folgende Begriffsverständnisse vor:

* Ein *Werturteil* ist das Ergebnis einer kognitiven Operation, bei der einem Gegenstand der Status eines *zu erstrebenden Gutes* (positives Gut) bzw. eines *zu vermeidenden Übels* (negatives Gut) zugeschrieben wird. Ein positives Gut kann auch als *Wert* bezeichnet werden[4]. Als typische Gegenstände kommen im Fall der Wissenschaft insbesondere die Eigenschaften von Theorien, Hypothesen, Methoden oder Modellen infrage. Werturteile lassen sich durch die *Art* des in ihnen gekennzeichneten Guts spezifizieren. Hier stellt sich im Fall der Wissenschaft insbesondere die Frage, ob ein Gut hinsichtlich der wissenschaftlichen Erkenntnis selbst (*epistemischer* Wert) oder einer anderen menschlichen Praxis (*nicht-epistemischer* Wert) erstrebenswert ist. Weiterhin lassen sich Werturteile danach unterscheiden, ob sie das Erstreben bzw. Vermeiden eines positiven bzw. negativen Guts lediglich *empfehlen* oder verpflichtend *vorschreiben*.
* *Wertfreiheit* bezeichnet einen Zustand einer menschlichen Praxis, in dem die an der Praxis teilnehmenden Personen Werturteile weder explizit fällen noch implizit voraussetzen[5]. Eine Praxis ist *vollständig* wertfrei, wenn alle Arten von Werturteilen

Wissenschaftskommunikation vorgelegt (John, 2018; John, 2019). Weiterhin werden Output- und Inputthemen häufig gemeinsam abgehandelt (z.B. Kourany, 2003; Douglas, 2009; Kitcher, 2011).

4 Zu beachten ist dabei, dass das Wort „Wert" im allgemeinen wie im philosophischen Sprachgebrauch für zwei zusammenhängende, konzeptuell aber zu unterscheidende Sachverhalte verwendet wird: zum einen dafür, dass ein Gegenstand erstrebenswert ist (etwa: „Freiheit ist ein hoher Wert"), zum anderen dafür, dass Personen bestimmte Werturteile für verbindlich halten (etwa: „Die Werte des Westens"). Im ersten Fall bezeichnet das Wort „Wert" den normativen Status eines Gegenstands, im zweiten Fall das Für-gültig-Halten des Werturteils, das dem Gegenstand einen solchen Status zuweist. In der obigen Definition wird „Wert" im ersteren Sinn verwandt.

5 Der Zusatz „implizit voraussetzen" ist wichtig, da Entscheidungen im Forschungsprozess häufig aufgrund disziplinärer Standards (Wilholt, 2009; Wilholt, 2013) und anderer Konventionen gefällt werden. Beispiele hierfür

aus allen Bereichen der Praxis ausgeschlossen sind. *Partiell* wertfrei ist eine Praxis hingegen, wenn lediglich bestimmte Bereiche der Praxis frei von Werturteilen sind (*kontextuelle* Wertfreiheit) oder lediglich bestimmte Arten von Werturteilen ausgeschlossen sind (*differenzielle* Wertfreiheit).

* *Wertbeladenheit* bezeichnet einen Zustand einer menschlichen Praxis, in dem Wertfreiheit entweder grundsätzlich nicht realisierbar ist (*prinzipielle* Wertbeladenheit) oder aus unsystematischen, nicht in der Natur der Praxis liegenden Gründen zu einem gegebenen Zeitpunkt nicht realisiert ist (*kontingente* Wertbeladenheit). Liegt Wertbeladenheit vor, so kann diese entweder einen maßgeblichen Einfluss auf die Praxis haben, etwa durch die Veränderung von Forschungsergebnissen in der Wissenschaft (*signifikante* Wertbeladenheit), oder aber in ihren Auswirkungen geringfügig oder beherrschbar sein (*insignifikante* Wertbeladenheit).
* Das *Wertfreiheitsideal* bezieht sich auf die Praxis der wissenschaftlichen Erkenntnis. Es geht davon aus, dass neben den offensichtlich wertbeladenen Bereichen dieser Praxis (Themenwahl, Anwendung, Forschungsethik) ein epistemischer Kernbereich existiert, der die spezifische Erkenntnisleistung von Wissenschaft konstituiert. Das Wertfreiheitsideal bezieht sich ausschließlich auf wissenschaftliche Aktivitäten in diesem Kernbereich (z.B. Datensammlung, Modellwahl, Hypothesenprüfung). Weiterhin bezieht es sich ausschließlich auf nicht-epistemische (z.B. ethische, politische, soziale) Werturteile. Eine mögliche Formulierung lautet:

  Innerhalb des epistemischen Kernbereichs von Wissenschaft ist es sowohl prinzipiell möglich als auch im verpflichtenden Sinn erstrebenswert, einen Zustand differenzieller und kontextueller Wertfreiheit herbeizuführen, in dem Wissenschaftlerinnen und Wissenschaftler nicht-epistemische Werturteile nicht in signifikantem Umfang fällen oder voraussetzen.

Das so verstandene Wertfreiheitsideal stellt eine komplexe, aus mehreren Teilthesen bestehende philosophische Position dar. Da diese für die wissenschaftsphilosophische Debatte von zentraler Bedeutung sind, ist es sinnvoll, die wichtigsten Aspekte des Wertfreiheitsideals separat hervorzuheben. Hierbei lässt sich zwischen einer normativen These (WFI_{Norm}), einer deskriptiven These (WFI_{Desk}), einer Kontextualitätsthese (WFI_{Kont}) und einer Differenzialitätsthese (WFI_{Diff}) unterscheiden[6]:

sind Signifikanzlevels in statistischen Tests (Douglas, 2000) oder Diskontraten in klimaökonomischen Modellen (Frisch, 2013). Die in diesen Konventionen enthaltenen Annahmen werden von individuellen Wissenschaftlerinnen und Wissenschaftlern typischerweise nicht explizit getroffen, sondern schlicht vorausgesetzt.

6 An anderer Stelle (s. Teil III dieser Arbeit) rekonstruiere ich das Wertfreiheitsideal in leicht modifizierter

WFI_{Norm}	Kontextuelle und differenzielle Wertfreiheit *soll* – im Sinne einer alle Wissenschaftlerinnen und Wissenschaftler betreffenden Verpflichtung – realisiert werden.
WFI_{Desk}	Kontextuelle und differenzielle Wertfreiheit *kann* – zumindest im Prinzip – in signifikantem Umfang realisiert werden.
WFI_{Kont}	Epistemische und nicht-epistemische *Bereiche* sind hinreichend unterscheidbar.
WFI_{Diff}	Epistemische und nicht-epistemische *Werturteile* sind hinreichend unterscheidbar.

Dieses Begriffsverständnis reflektiert weite Teile der jüngeren und klassischen Wertfreiheitsdebatte[7] (z.B. Weber, 1904/1988; Reichenbach, 1938/1961; Rudner, 1953; Popper, 1974; Kuhn, 1977; McMullin, 1982; Lacey, 1999; Dorato, 2004; Douglas, 2009; Betz, 2013; Biddle, 2013; Bueter, 2015; Reiss & Sprenger, 2020). Bevor ich die einzelnen Elemente des Wertfreiheitsideals diskutiere und in der Literatur verorte, möchte ich auf das Verhältnis der Teilthesen untereinander eingehen.

2.3 WERTFREIHEIT: VERHÄLTNIS DER TEILTHESEN

Bei der Betrachtung der Teilthesen des Wertfreiheitsideals fällt auf, dass diese zwei verschiedene Gruppen bilden: Während WFI_{Norm} und WFI_{Desk} die in normativer und deskriptiver Dimension erhobenen Geltungs*ansprüche* des Wertfreiheitsideals beschreiben – also die Hinsichten, in denen die Frage, *ob* das Wertfreiheitsideal gilt, sinnvoll gestellt ist – beziehen sich WFI_{Kont} und WFI_{Diff} dessen Geltungs*bereich*. Genauer gesagt benennen WFI_{Kont} und WFI_{Diff} die Bedingungen, unter denen der Geltungsbereich des Wertfreiheitsideals wohldefiniert ist. Das Wertfreiheitsideal als philosophische Position zu

Form, nämlich als zwei Hauptthesen und vier Einschränkungen. Die Rekonstruktion im vorliegenden Teil der Arbeit unterscheidet sich insofern hiervon, als sie die Einschränkungen auf *prinzipielle* und *signifikante* Wertfreiheit als Teil von WFI_{Desk} und die Einschränkungen auf *nicht-epistemische* Werturteile und den *epistemischen Kernbereich* als selbständige Thesen behandelt.

7 Dennoch lassen sich, wie bei Begriffsdefinitionen dieser Art üblich, auch andere Verständnisse denken. So werden etwa epistemische Werturteile manchmal auch als kognitive Werturteile bezeichnet (Longino, 1996; Lacey, 1999); andererseits werden kognitive und epistemische Werturteile manchmal nicht synonym, sondern in Abgrenzung zueinander definiert (Laudan, 2004; Douglas, 2009, Kap. 5). Darüber hinaus werden die deskriptive These der Möglichkeit von Wertfreiheit und die normative These ihrer Erwünschtheit manchmal als zwei verschiedene Behauptungen (Reiss & Sprenger, 2020) und manchmal als Aspekte einer einzigen philosophischen Position betrachtet (Biddle, 2013).

vertreten bedeutet dann, das Gelingen der wissenschaftlichen Erkenntnispraxis von der Realisierung eines Zustandes abhängig zu machen, der qua WFI_{Norm} als erstrebenswert und qua WFI_{Desk} als de facto realisierbar betrachtet wird, und zwar innerhalb des von WFI_{Kont} und WFI_{Diff} als wohldefiniert behaupteten Geltungsbereichs. Dieser Geltungsbereich spielt eine wichtige Rolle, da einige Kritiken an WFI_{Desk} auf einem Angriff auf WFI_{Kont} oder WFI_{Diff} basieren: Kann gezeigt werden, dass die Grenzen des Wertfreiheitsideals nicht wohldefiniert sind, so kann nicht behauptet werden, dass Wertfreiheit innerhalb dieser Grenzen möglich ist (z.B. Longino, 1996; Machamer & Douglas, 1999; Putnam, 2002; Dupré, 2007; Bueter, 2015; De Melo-Martin & Intemann, 2016). Andere Kritiken attackieren WFI_{Desk} direkt, also ohne den Umweg über WFI_{Kont} oder WFI_{Diff}. Nach diesen direkten Kritiken ist WFI_{Desk} selbst dann ungültig, wenn ausschließlich genuin wissenschaftliche Aktivitäten im epistemischen Kernbereich betrachtet werden (z.B. Rudner, 1953; Wilholt, 2009).

Mit Blick auf WFI_{Norm} und WFI_{Desk} ist weiterhin beachtenswert, dass die beiden Geltungsansprüche jeweils andere Arten der Kritik eröffnen: Ein Angriff auf WFI_{Desk} kann entweder direkt oder über den Umweg von WFI_{Kont} oder WFI_{Diff} erfolgen. Da die normative Qualität von Wertfreiheit keine Auswirkungen auf ihre Möglichkeit hat, kann WFI_{Desk} jedoch nicht über WFI_{Norm} kritisiert werden. In entgegengesetzter Richtung verhält es sich anders, da Sollsätze mit deskriptiven Mitteln kritisiert werden können, wenn ein entsprechendes Argument zusätzlich normative Annahmen – etwa ein metaethisches Prinzip wie Sollen-impliziert-Können – enthält. Neben einem direkten Angriff auf WFI_{Norm} besteht daher die Möglichkeit, den normativen Geltungsanspruch des Wertfreiheitsideals indirekt über WFI_{Desk} zu kritisieren. Gleiches gilt für den indirekten Angriff auf WFI_{Norm} über WFI_{Kont} und WFI_{Diff}. In diesem Fall müsste zunächst mit deskriptiven Mitteln das Scheitern von WFI_{Kont} und/oder von WFI_{Diff} gezeigt werden, um dann unter Zuhilfenahme normativer Prinzipien wie Sollen-impliziert-Können eine Zurückweisung von WFI_{Norm} zu erreichen. Abbildung 1 veranschaulicht die genannten Kritikoptionen gemeinsam mit einer Kurzform des Wertfreiheitideals und seiner Teilthesen.

Nächste Seite:
Abbildung 1: Inhalt, Teilthesen und mögliche Kritiken des Wertfreiheitsideals.

Inhalt des Wertefreiheitsideals

Innerhalb des epistemischen Kernbereichs ist es prinzipiell möglich und im verpflichtenden Sinne erstrebenswert, nicht-epistemische Werturteile nicht in signifikantem Umfang zu fällen oder vorauszusetzen.

Teilthesen des Wertfreiheitsideals

WFI_{Norm}	Kontextuell-differenzielle Wertfreiheit *soll* realisiert werden (im verpflichtenden Sinne).	Geltungsansprüche in normativer und deskriptiver Dimension
WFI_{Desk}	Kontextuell-differenzielle Wertfreiheit *kann* in signifikantem Umfang realisiert werden (im Prinzip).	
WFI_{Kont}	Epistemische und nicht-epistemische *Wissenschaftsbereiche* sind hinreichend unterscheidbar.	Geltungsbereich nach Wissenschaftsbereichen und Werturteilsarten
WFI_{Diff}	Epistemische und nicht-epistemische *Werturteile* sind hinreichend unterscheidbar.	

Mögliche Kritiken des Wertfreiheitsideals

Kritisierbar über:	WFI_{Norm}	WFI_{Desk}
Via WFI_{Norm}	Ja, direkte Kritik mit *normativen* Argumenten	Nein
Via WFI_{Desk}	Ja, indirekte Kritik mit *deskriptiven* und *normativen* Argumenten	Ja, direkte Kritik mit *deskriptiven* Argumenten
Via WFI_{Kont}	Ja, indirekte Kritik mit *deskriptiven* und *normativen* Argumenten	Ja, indirekte Kritik mit *deskriptiven* Argumenten
Via WFI_{Diff}	Ja, indirekte Kritik mit *deskriptiven* und *normativen* Argumenten	Ja, indirekte Kritik mit *deskriptiven* Argumenten

2.4 WERTFREIHEIT – NUR EIN IDEAL?

Angesichts dieser Angriffsvektoren stellt sich nun aber die Frage, ob Unterstützerinnen und Unterstützer des Wertfreiheitsideals tatsächlich WFI_{Desk} vertreten müssen. Denn während die Funktion von WFI_{Kont} und WFI_{Diff} darin besteht, das Wertfreiheitsideal *einzuschränken* (auf kontextuelle und differenzielle Wertfreiheit, s. Kap. 5), läuft WFI_{Desk} auf eine *Erweiterung* im Vergleich zu einem rein normativen Verständnis hinaus. Da die Kombination von WFI_{Norm} und WFI_{Desk} offenbar schwerer zu verteidigen ist als WFI_{Norm} alleine, kann es aus Sicht von Vertreterinnen und Vertretern des Wertfreiheitsideals sinnvoller erscheinen, dieses als *bloßes* Ideal, also als Handlungsregulativ ohne Möglichkeitsanspruch zu begreifen. Nach dieser Lesart ist das Wertfreiheitsideal "just that: an ideal" (Bright, 2018, S. 2244). Auf den ersten Blick scheinen einige Beispiele aus der Literatur in diese Richtung zu weisen: „defenders of the so-called value-freedom or value-neutrality have argued again and again that science is not possible without […] values" (Zecha, 1992, S. 155). In diesem Sinne beschreibt etwa Popper Wertfreiheit als Ideal, „das vermutlich nie erreichbar ist" (1974, S. 114). Weiter schreibt Popper: „Ohne Leidenschaft geht es nicht, und schon gar nicht in der reinen Wissenschaft. Das Wort ‚Wahrheits*liebe*' ist keine bloße Metapher" (ebd., Hervorh. i. Orig.). Auch Weber betont die Unmöglichkeit vollständiger Wertfreiheit:

> Es gibt keine schlechthin ‚objektive' wissenschaftliche Analyse […] der ‚sozialen Erscheinungen' *unabhängig* von speziellen und ‚einseitigen' Gesichtspunkten, nach denen sie […] als Forschungsobjekt ausgewählt, analysiert und darstellend gegliedert werden (Weber, 1904/1988, S. 170, Hervorh. i. Orig.).

Warum ist es dennoch gerechtfertigt, das Wertfreiheitsideal als normativen *und* deskriptiven Geltungsanspruch zu rekonstruieren? Ich sehe hierfür zwei Gründe. Der erste Grund besteht darin, dass viele der einschlägigen Verteidigungen des Wertfreiheitsideals keineswegs im Widerspruch zu WFI_{Desk} stehen, und in einigen Fällen sogar explizit WFI_{Desk} oder eine verwandte These vertreten (Justin Biddle zeigt dies etwa für Richard Jeffrey und Sandra Mitchell, s. Biddle, 2013, S. 131). Denn die dort geäußerten Zweifel an der Möglichkeit von Wertfreiheit beziehen sich häufig auf *vollständige*, nicht aber die von WFI_{Desk} adressierte *partielle* Wertfreiheit. So handelt es sich etwa bei der von Popper erwähnten Wahrheitsliebe um ein epistemisches Werturteil, während sich Weber auf Werturteile jenseits des epistemischen Kernbereichs, nämlich die Auswahl und Operationalisierung von Forschungsproblemen bezieht. Weiterhin kann Wertfreiheit selbst dann, wenn es „*praktisch* unmöglich [ist], die außerwissenschaftlichen Werte aus dem

Wissenschaftsbetrieb zu verbannen“ (Popper, 1974, S. 114, meine Hervorh.), dennoch *im Prinzip* möglich sein. In diesem Sinne verstehen viele Vertreterinnen und Vertreter des Wertfreiheitsideals die Wertbeladenheit einzelner Forschungsergebnisse als zwar verbreitetes, aber *transientes* Phänomen (Ruphy, 2006; Koertge, 2013). So argumentiert etwa Popper: „Solche Kleinigkeiten wie zum Beispiel der soziale oder ideologische Standort des Forschers schalten sich […] mit der Zeit von selber aus, obwohl sie natürlich kurzfristig immer eine Rolle spielen“ (Popper, 1974, S. 113). Ähnlich argumentiert Ernan McMullin:

> Other scientists attempt to duplicate experimental claims; theoreticians try to extend the theories involved in new and untried ways […] and so on. […] To the extent that non-epistemic values and other non-epistemic factors have been instrumental in the original theory-decision […], they are gradually sifted […]. The non-epistemic, by very definition, will not in the long run survive this process (McMullin, 1982, S. 23).

Dass Wertbeladenheit ein transientes Phänomen ist, heißt nun aber nichts anderes, als dass Wertfreiheit prinzipiell möglich ist. Denn damit sich Wertfreiheit auf lange Sicht durchsetzen kann, müssen Wissenschaftlerinnen und Wissenschaftler in der Lage sein, die Werteinflüsse existierender Forschungsergebnisse zunächst zu erkennen, dann in den eigenen Studien zu eliminieren und dies so lange fortsetzen, bis schließlich „die Reinheit der reinen Wissenschaft“ (Popper, 1974, S. 114) in Bezug auf die vorliegende Forschungsfrage hinreichend verwirklicht ist. Die von Befürworterinnen und Befürwortern des Wertfreiheitsideals häufig betonte *de facto* Wertbeladenheit von Wissenschaft steht somit keineswegs im Widerspruch zu $\mathrm{WFI}_{\mathrm{Desk}}$.

Der zweite Grund für die Integration von $\mathrm{WFI}_{\mathrm{Desk}}$ in das Verständnis des Wertfreiheitsideals liegt in dem bereits erwähnten metaethischen Grundsatz *Sollen-impliziert-Können*. Nach diesem Grundsatz kann ein Gebot, das Anspruch auf Gültigkeit erhebt, nicht prinzipiell unerfüllbar sein. Folglich erscheint es unplausibel, Wertfreiheit qua $\mathrm{WFI}_{\mathrm{Norm}}$ als verpflichtend zu postulieren, ohne gleichzeitig qua $\mathrm{WFI}_{\mathrm{Desk}}$ ihre Möglichkeit zu behaupten. Das zugrunde liegende Prinzip wird häufig auf Kant zurückgeführt, findet sich aber bereits in der römischen Rechtsprechung als *impossibilium nulla obligatio est* (Digesta 50, 17, 185). Ich werde an dieser Stelle nicht auf alle mit Sollen-impliziert-Können verknüpften Lesarten eingehen (eine gute Übersicht bietet Kühler, 2013), sondern nur kurz einen naheliegenden Einwand diskutieren und das Prinzip ansonsten schlicht voraussetzen. Dies scheint durch dessen intuitive Plausibilität als „commonplace of practical reasoning“ (Jacobs, 1985, S. 43) sowie durch die Tatsache gerechtfertigt, dass selbst

Kritikerinnen und Kritiker das Prinzip meist in irgendeiner Form anerkennen. Kontrovers ist daher weniger die Gültigkeit von Sollen-impliziert-Können als vielmehr seine Interpretation: „while few would reject the principle altogether, there is disagreement about how exactly it should be understood" (Stern, 2004, S. 43)[8].

Ein naheliegender Einwand gegen die Anwendung von Sollen-impliziert-Können auf das Wertfreiheitsideal entspringt einer dieser Interpretationen. So favorisiert etwa Robert Stern eine schwache Lesart, nach der das Prinzip für Pflichten, nicht aber für Ideale gilt. So könne man etwa – kontrafaktisch – sagen: "it would be better if the child *could* read Shakespeare rather than just Harry Potter books" (Stern, 2004, S. 51, Hervorh. i. Orig.). Auch andere Autorinnen und Autoren verstehen Ideale in diesem nicht verpflichtenden Sinn. Als Beispiel nennt etwa Brian Talbott „the ideal to love all people as one loves oneself" (Talbott, 2016, S. 380). Derartige Ideale, so das Argument, sind unrealisierbar und verfügen doch über normative Kraft. Ähnlich argumentiert Max Weber: So wie das Sittengesetz „unerfüllbar ist, dennoch aber als ‚aufgegeben' gilt" (Weber, 1917/1988, S. 497), so sei Wertfreiheit auch dann erstrebenswert, wenn sie praktisch nicht oder kaum realisierbar ist.

Ist es somit falsch, die Integration von WFI_{Desk} in das Wertfreiheitsideal mit Sollen-impliziert-Können zu begründen? Ich denke, dass dieser Eindruck täuscht, und zwar aus den bereits genannten Gründen: Praktische Schwierigkeiten bei der Verwirklichung von Wertfreiheit implizieren solange keine *prinzipielle* Unmöglichkeit, und stehen somit nicht im Widerspruch zu WFI_{Desk}, wie die begründete Aussicht besteht, die verhindernden Faktoren durch (gegebenenfalls radikale oder erst nachträglich wirksame) Maßnahmen auszuräumen oder zu kompensieren. Ideale wie universelle Liebe, das Lesen von

8 Sollen-impliziert-Können scheint auf dem ersten Blick einem anderen Prinzip zu widersprechen, das ich ebenfalls in dieser Arbeit diskutiere: Kein-Sein-aus-Sollen (Teil I und Teil III). Dieser Eindruck ist jedoch irreführend. Denn während es sich im Fall von Kein-Sein-aus-Sollen um ein *logisches* Prinzip handelt, beschreibt Sollen-impliziert-Können den *ethischen* Grundsatz, dass verpflichtende Normen erfüllbar sein sollen. Dieser Grundsatz ist jedoch nicht so zu verstehen, dass aus einer normativen Prämisse (dem Sollen) eine deskriptive Konklusion (das Können) *erschlossen* wird. Vielmehr geht es darum, dass die normative Prämisse ethisch inakzeptabel ist, wenn die deskriptive Konklusion falsch ist. Anders als bei einem fehlschlüssigen Sollen-Sein-Schluss werden somit keine deskriptiven Gehalte aus dem Nichts erzeugt. Ebenso wird die Schlussbarriere nicht in die andere Richtung verletzt, wenn Sollen-impliziert-Können zur Zurückweisung einer unerfüllbaren Norm genutzt wird. Denn auch in diesem Fall stammt die entsprechende normative Konklusion (die Ungültigkeit der Norm) nicht aus einer deskriptiven Prämisse (der Unerfüllbarkeit der Norm), sondern aus der zusätzlichen normativen Prämisse, dass Normen erfüllbar sein *sollen*.

Shakespeare durch Kinder oder ein Leben nach dem kategorischen Imperativ sind daher nicht *prinzipiell* unmöglich, sondern lediglich *sehr schwer* umzusetzen. Ausnahmefälle – Kinder mit Dyslexie (Shakespeare), Personen mit bestimmten neurologischen Krankheiten (Liebe), moralische Dilemmata (Sittengesetz) – sind nicht schlagend gegen die Anwendung von Sollen-impliziert-Können, da es ebenso plausibel ist, diese Fälle von der Geltung des jeweiligen Ideals auszunehmen.

Damit ist jedoch nicht behauptet, dass es überhaupt keine Ideale gibt, die unrealisierbar und dennoch erstrebenswert sind (etwa bestimmte religiöse Ideale). Dies ist jedoch weder ein Argument gegen Sollen-impliziert-Können noch dafür, Wertfreiheit als ein Ideal dieser Art zu begreifen. Denn wenn unrealisierbare Ideale Verpflichtungen erzeugen, dann nicht zur Realisierung, sondern zur Approximation – und da Approximationen typischerweise möglich sind, wird Sollen-impliziert-Können hierdurch nicht verletzt. Wäre Wertfreiheit also ein prinzipiell unrealisierbares Ideal, so könnte $\mathrm{WFI}_{\mathrm{Norm}}$ nicht die *Realisierung von Wertfreiheit* fordern, sondern höchstens die *Reduzierung von Wertbeladenheit* (s. ähnlich Biddle, 2013, S. 131; Kitcher, 2011, S. 39–40). Doch auch in diesem Fall würde das Wertfreiheitsideal darauf hinaus laufen, dass zumindest *einige* Werturteile – nämlich diejenigen, die qua Approximation eliminierbar sind – nicht gefällt werden dürfen und dass dies auch tatsächlich möglich ist. Eine approximative Interpretation berechtigt somit höchstens zu einer schwächeren Lesart des Wertfreiheitsideals, nicht aber zum völligen Rückzug aus der deskriptiven Geltungsdimension. Ich werde auf die schwächere Lesart und die damit zusammenhängende Logik der Approximation im folgenden Kapitel zurückkommen. An dieser Stelle soll zunächst nur festgehalten werden, dass die Integration von $\mathrm{WFI}_{\mathrm{Desk}}$ in das Wertfreiheitsideal trotz der allgemein geteilten Auffassung der de facto bestehenden Wertbeladenheit von Forschung gerechtfertigt ist.

3. Die normative These ($\mathrm{WFI}_{\mathrm{Norm}}$)

3.1 DER TERMINUS „VERPFLICHTEND“

In diesem Kapitel werde ich die Diskussion von $\mathrm{WFI}_{\mathrm{Norm}}$, also des Sollensanspruchs des Wertfreiheitsideals, ein wenig vertiefen. Zwei Aspekte stehen dabei im Mittelpunkt: zum einen der Terminus „verpflichtend“ und das Problem schwer zu erfüllender Pflichten;

zum anderen das Problem, dass WFI_{Norm} seinerseits ein Werturteil ist. Ich beginne mit dem Aspekt der Verpflichtung.

Der in meiner Formulierung von WFI_{Norm} verwendete Terminus „verpflichtend" bezieht sich auf den Unterschied zwischen *Empfehlungen* und *Vorschriften*. Dies ist in zweifacher Hinsicht relevant für das Wertfreiheitsideal. Erstens beansprucht WFI_{Norm}, nicht nur die wissenschaftliche Praxis *als Ganzes*, sondern jede einzelne Wissenschaftlerin und jeden einzelnen Wissenschaftler normativ zu binden. Hierin unterscheidet es sich von kollektiven Idealen wie dem der distributiven Gerechtigkeit. So lässt sich etwa aus dem gesamtgesellschaftlichen Ziel der Armutsbekämpfung keine individuelle Pflicht zu karitativem Engagement (z.B. in Form umfangreicher Spenden) ableiten. Im Gegensatz dazu verpflichtet WFI_{Norm}, wenn es gültig ist, jede Wissenschaftlerin und jeden Wissenschaftler als Individuum. Dies wird auch nicht durch die erwähnte Position mancher Autorinnen und Autoren infrage gestellt, nach der Wertfreiheit erst durch den fortgesetzten wissenschaftlichen Diskurs möglich wird (Popper, 1974; McMullin, 1982). Denn diese Position ist nicht so zu verstehen, dass es einerseits Aufgabe „der wissenschaftlichen Diskussion [ist], die Vermengung der Wertsphären zu bekämpfen" (Popper, 1974, S. 114), während andererseits auf individueller Ebene keine Wertfreiheit angestrebt werden muss. Vielmehr sollen individuelle Wissenschaftlerinnen und Wissenschaftler zur Verwirklichung des Ideals beitragen, selbst wenn dies nicht immer gelingt.

Die zweite Hinsicht, in der WFI_{Norm} einen Anspruch auf Verpflichtung erhebt, hängt eng damit zusammen. Hierbei geht es jedoch weniger um den Unterschied zwischen individueller und kollektiver Geltungsebene, sondern um den zwischen *gesollten* und *supererogatorischen* Handlungen. Denn offensichtlich kann ein Gut erstrebenswert sein, ohne dass hieraus eine Pflicht zur Realisierung dieses Gutes entspringt. So existiert etwa keine Pflicht zur Erlangung einer hohen Bildung, obwohl Bildung zweifellos ein Ideal ist, das sich auf das Individuum bezieht. Das Wertfreiheitsideal ist nicht von dieser Art. Vielmehr stellt es die anspruchsvolle Forderung, Wertfreiheit *neigungsunabhängig* und insbesondere auch dann umzusetzen, wenn hierfür Kosten (etwa ein höherer Forschungsaufwand) in Kauf genommen werden müssen. In diesem Sinne betont Max Weber, dass es nicht darum gehen könne, „ob die Scheidung von empirischer Feststellung und praktischer Wertung ‚schwierig' sei" (Weber, 1917/1988, S. 497) – schließlich, so Weber, sei die Antwort hierauf klar: „Sie ist es" (ebd.). Da WFI_{Norm} keinen bloß supererogatorischen, sondern einen verpflichtenden Anspruch erhebt, können Realisierungsschwierigkeiten nicht, wie im Falle des Bildungsideals, ohne weiteres das Aufgeben des

Ideals begründen. Wenn WFI_{Norm} gültig ist, gilt vielmehr: „Es mag sein, dass wir [bei der Realisierung von Wertfreiheit] nicht immer erfolgreich sind, aber das bedeutet nur, dass wir uns mehr anstrengen sollten, nicht dass wir den Versuch aufgeben sollten" (Koertge, 2013, S. 247).

Der Unterscheid zwischen kollektiven und individuellen Idealen sowie zwischen supererogatorischen und gesollten Handlungen lässt allerdings offen, worin genau die *Erfüllungsbedingungen* von WFI_{Norm} bestehen: Ist die Verpflichtung erfüllt, wenn alle verfügbaren und zumutbaren Mittel (methodische Strenge, Einholen fachlicher Kritik etc.) ausgeschöpft wurden? Oder muss Wertfreiheit hierfür auch tatsächlich erreicht worden sein? In der Literatur finden sich für beide Interpretationen Belege. Die erste, schwächere Lesart entspricht dem im vorigen Kapitel erwähnten approximativen Verständnis des Wertfreiheitsideals. Hiernach kann WFI_{Norm} etwa wie folgt verstanden werden:

> Scientists should strive to minimize the influence of contextual values on scientific reasoning, e.g., in gathering evidence and assessing/accepting scientific theories (Reiss & Sprenger, 2020).
> [T]he value judgments internal to science, involving the evaluation and acceptance of scientific results […], are to be as free as humanly possible of all social and ethical values (Douglas, 2009, S. 45).

Die schwächere Lesart zeigt sich hier in Ausdrücken wie „strive to", „minimize" oder „as free as humanly possible". Auf dieser Grundlage ließe sich argumentieren, dass Wissenschaftlerinnen und Wissenschaftler WFI_{Norm} auch dann genügen können, wenn ihre Forschung zwar de facto wertbeladen ist, sie aber alles unternommen haben, um die Wertbeladenheit so weit wie möglich zu reduzieren. Demgegenüber existiert in der Literatur auch die stärkere Lesart, nach der die Erfüllungsbedingungen von WFI_{Norm} das tatsächliche Realisieren von Wertfreiheit beinhalten[9]. Kennzeichnend hierfür sind, neben dem Wegfall approximativer Begriffe, etwa Ausdrücke wie „only" und „all":

> Theory assessment has to be based on empirical evidence and cognitive values only (Bueter, 2015, S. 20).

9 Tatsächlich finden sich beide Lesarten teilweise an ein und derselben Stelle. So schreibt etwa Douglas direkt nach ihrer „as-free-humanly-possible"-Formulierung: "[the value-free ideal claims that] scientific judgments are to be driven by values wholly internal to the scientific community" (Douglas, 2009, S. 45). Offensichtlich ist es aber ein Unterschied, ob *ausschließlich* („wholly") epistemische Werturteile zugelassen sind, oder ob nicht-epistemische Werturteile lediglich *weitestmöglich* („as free as humanly possible") minimiert werden sollen.

> [S]cientists ought to apply scientific methods properly, thereby screening out all contextual factors (Biddle, 2013, S. 125).
> [T]he justification of scientific findings should not be based on non-epistemic (e.g. moral or political) values (Betz, 2013, S. 207).

Diese stärkere Lesart hat jedoch ein Problem: Wenn die Erfüllungsbedingungen von WFI_{Norm} das tatsächliche Erreichen von Wertfreiheit beinhalten, dann erscheinen große Teile des Forschungsbetriebs als tadelnswert. Schließlich ist Wertbeladenheit auch aus Sicht von Vertreterinnen und Vertretern des Wertfreiheitsideals weit verbreitet (Weber, 1917/1988; Popper, 1974; McMullin, 1982; Ruphy, 2006; Koertge, 2013). Selbst die herausragendsten Episoden der Wissenschaftsgeschichte sind hiervon nicht ausgenommen: „Historical scholarship […] has suggested that the work of even the greatest scientists – even scientists like Boyle, Darwin, and Freud, and even, perhaps, the great Newton and Einstein themselves – was shaped by social values" (Kourany, 2008, S. 88). Nun scheint aber ein Ideal, das selbst derart wichtige Beiträge zum wissenschaftlichen Fortschritt als tadelnswert darstellt, wenig überzeugend (Longino, 2004; Kourany, 2008). Dieses Problem ist zwar nicht unbedingt spezifisch für die stärkere Lesart; dennoch erscheint es hier besonders relevant, da ihre Erfüllungsbedingungen unrealistischer als die der schwächeren Lesart sind. Es ist daher kein Zufall, dass viele Vertreterinnen und Vertreter des Wertfreiheitsideals eher der schwächeren Lesart zuneigen (Weber, 1917/1988; Popper, 1974; McMullin, 1982; Ruphy, 2006; Bright, 2018).

Ist es somit sicher, dass WFI_{Norm} nicht plausibel in der stärkeren Lesart verstanden werden kann? Nicht unbedingt. Erstens impliziert die stärkere Lesart nicht automatisch, wie in der obigen Argumentation behauptet, eine weitreichende Ablehnung des Forschungsbetriebs. Denn offensichtlich ist der mit einer Pflichtverletzung verbundene moralische Tadel ein *graduelles* Phänomen – und zwar selbst dann, wenn die verletzte Pflicht auf vollständige Realisierung anstatt auf bloße Approximation abzielt. So kann etwa das Ideal der Rechtschaffenheit im starken Sinn interpretiert werden („sei rechtschaffend" anstatt „sei so rechtschaffend wie möglich"), ohne damit alle nicht rechtschaffenden Handlungen im gleichen Maße abzulehnen[10]. Die moralische Bewertung

10 Der folgende Fall verdeutlicht dies beispielhaft: Eine Person nimmt eine kostenpflichtige Leistung in Anspruch; der fragliche Dienstleister versäumt die Rechnungstellung; die Person weist den Dienstleister hierauf hin; nachdem dieser Hinweis ohne Reaktion bleibt, behält die Person den schuldigen Betrag ein. Offensichtlich nimmt die Person ihre Rechtschaffenheitspflichten nicht vollumfänglich war (sie hätte die Rechnung ein weiteres Mal einfordern können). Dennoch begründet dies keinen schweren moralischen Tadel. Insbesondere begründet es keinen Tadel, der mit schwerwiegenden Verletzungen der Rechtschaffenheitspflicht (etwa großangelegtem Betrug) vergleichbar wäre.

einer Handlung hängt nicht nur davon ab, *ob* eine Pflicht verletzt wurde, sondern auch davon, welche *Folgen* die Pflichtverletzung hatte, welche *Motive* beteiligt waren und ob zumindest der *Versuch* der Pflichterfüllung unternommen wurde. Folglich ist es denkbar, nicht erreichte Wertfreiheit als Pflichtverletzung zu deuten, diese Pflichtverletzung aber unter bestimmten Umständen zu *entschuldigen* (z.B. wenn Forschung trotz intensiver Bemühungen und ohne das Wissen der beteiligten Personen wertbeladen ist). Die entsprechende Forschung wäre zwar aus Sicht von $\mathrm{WFI}_{\mathrm{Norm}}$ weiterhin defizitär; jedoch erzwingt dies nicht die unplausible Behauptung, dass ein Großteil der zeitgenössischen und historischen Wissenschaft abzulehnen ist.

Der zweite Grund, der die stärkere Lesart stützt, liegt in einer impliziten Annahme der schwächeren Lesart. Dieser liegt offenbar die These zugrunde, dass jeder Zustand S jedem anderen Zustand S^* vorzuziehen ist, wenn S das Wertfreiheitsideal stärker approximiert als S^*. Dies ist jedoch keineswegs trivial. Tatsächlich gibt es Fälle, in denen nicht realisierbare Handlungsziele besser aufgegeben als approximiert werden sollten. So sind etwa für einen Gestrandeten, der von einer Insel zum rettenden Festland schwimmen möchte, die Etappenziele (das offene Meer) nur dann dem Ausgangspunkt (der Insel) vorzuziehen, wenn das Endziel (das Festland) tatsächlich erreichbar ist. Ist dies nicht gegeben, sind Zielapproximationen nicht erstrebenswert (für eine klassische Darstellung dieses als *Problem of the Second-Best* bekannten Themas siehe Lipsey & Lancaster, 1956)[11].

Auf ähnliche Weise ließe sich argumentieren, dass Wertfreiheit, wenn sie nicht vollständig realisierbar ist, auch nicht annäherungsweise realisiert werden soll. Da nicht-epistemische Werturteile häufig mit sozialen Perspektiven und Gruppen assoziiert sind, ist es zum Beispiel denkbar, dass approximative Wertfreiheit die Sichtweisen bestimmter – aber nicht aller – Gruppen eliminiert, so dass die verbleibenden sozialen Sichtweisen unangemessen dominant werden. Es wäre sogar denkbar, dass *mehr* Wertbeladenheit zu *weniger* Verzerrung führt, etwa durch gegenseitige Ausbalancierung der Perspektiven (Longino 1990; 2002). Weiterhin könnten diese Effekte unsystematisch sein, so dass approximative Wertfreiheit manchmal zu geringerer, manchmal aber zu größerer Verzerrung führt. In diesem Fall müsste $\mathrm{WFI}_{\mathrm{Norm}}$ in der schwächeren Lesart mit einem Einzelfallvorbehalt versehen werden, wodurch das Wertfreiheitsideal seinen universellen Anspruch einbüßen würde.

11 In Teil IV dieser Arbeit greife ich dieses Problem noch einmal im Zusammenhang mit Philip Kitchers (2011) Ideal der Deliberation auf.

Ich werde an dieser Stelle keine Position dazu beziehen, ob diese Argumente tatsächlich erfolgreich sind. In der Tat verwende ich selbst an anderer Stelle (s. Teil III) eine Formulierung des Wertfreiheitsideals, die eher der schwächeren Lesart ähnelt (allerdings hat dies keine Auswirkungen auf die dort vertretene Thesen). Mein Punkt ist hier lediglich, dass approximative Zustände nicht, wie in der schwächeren Lesart vorausgesetzt, generell erstrebenswert seien müssen. Entgegen dem ersten Eindruck ist daher die stärkere Lesart nicht unbedingt unplausibler als schwächere Lesart, was sie für Vertreterinnen und Vertretern des Wertfreiheitsideals durchaus attraktiv macht[12].

3.2 DER WERT VON WERTFREIHEIT – EIN WIDERSPRUCH?

Mit WFI_{Norm} wird der Gegenstand „Wertfreiheit" als positives Gut gekennzeichnet. Einem Gegenstand den Status eines positiven Gutes zuzuweisen bedeutet aber, ein Werturteil zu fällen. Somit gilt, worauf in der Literatur bereits vielfach hingewiesen wurde: „Auch die Wertfreiheit ist ein Wert" (Luhmann, 1971, S. 255). Hieraus scheint sich nun ein Problem zu ergeben, denn der Inhalt des mit WFI_{Norm} gefällten Werturteils besteht ja gerade darin, dass Werturteile *nicht* gefällt werden sollen. In diesem Sinne stellt etwa Popper fest: „da Wertfreiheit […] selbst ein Wert ist, ist die Forderung der unbedingten Wertfreiheit ein Paradox" (Popper, 1974, S. 115). Dieselbe Überlegung findet sich auch in neueren Beiträgen: „value freedom is a logical and performative contradiction, because value freedom […] is itself a value" (Letherby et al., 2013, S. 59). Wie entsteht dieser Eindruck der Selbstwidersprüchlichkeit und was folgt hieraus für WFI_{Norm}?

In Anlehnung an die bekannte Unterscheidung zwischen homo- und heterologischen Prädikaten (Grelling & Nelson, 1908) kann man zwischen homo- und heterologischen Handlungsregeln unterscheiden[13]. Eine Regel ist *homologisch*, wenn sie innerhalb ihres

12 Hierauf könnte erwidert werden, dass dasselbe Argument auch gegen die stärkere Lesart gerichtet werden kann. In der Tat können nicht-intendierte Folgen WFI_{Norm} unabhängig von der Lesart infrage stellen. Dies trifft aber nicht den vorgestellten Punkt. Denn im Gegensatz zur schwächeren Lesart betrachtet die stärkere Lesart approximative Zustände nicht als generell erstrebenswert. Während somit die schwächere Lesart einen wertfreieren Zustände auch dann einem wertbeladeneren Zustand vorziehen muss, wenn der wertfreiere Zustand normativ problematischer ist als der wertbeladenere Zustand, kann die stärkere Lesart beide Zustände gleichermaßen ablehnen.

13 Kurt Grelling und Leonard Nelson verwendeten die Unterscheidung, um damit eine Variante von Russels Mengenantinomie zu formulieren. Im Unterschied zu Grelling und Nelson geht es mir nicht um die paradoxale Struktur des Prädikats „heterologisch" (das Prädikat ist heterologisch, wenn es homologisch ist und homologisch, wenn es heterologisch ist), sondern um den Gegenbegriff „homologisch". Meine Diskussion läuft darauf hinaus,

eigenen Geltungsbereichs liegt, d.h. wenn die in ihr aufgestellten Gebote und Verbote für die Regel selbst gelten. Aufgrund dieser Selbstanwendbarkeit sind homologische Regeln, die ihren eigenen Forderungen nicht entsprechen, selbstwidersprüchlich. Beispiele wären ein Diskriminierungsverbot, das diskriminierende Ausdrücke enthält (semantische Inkonsistenz), oder ein in unfreundlichem Ton vorgetragenes Freundlichkeitsgebot (performative Inkonsistenz). *Heterologische* Regeln hingegen liegen außerhalb ihres eigenen Geltungsbereichs, d.h. diese Regeln sind nicht auf sich selbst anwendbar. Hierzu zählen etwa Verkehrs- oder Spielregeln, da diese ihren Gegenstand (den Verkehr, ein Spiel) regulieren, nicht aber sich selbst. Das Problem der Selbstwidersprüchlichkeit ergibt sich folglich nur für homologische Regeln. Übertragen auf das Wertfreiheitsideal bedeutet dies, dass der Eindruck der Selbstwidersprüchlichkeit offenbar aus der Interpretation von $\mathrm{WFI}_{\mathrm{Norm}}$ als homologische Handlungsregel entspringt. Dieser Widerspruch könnte semantischer Art sein – die Regel beinhaltet Ausdrücke („soll"), die sie verbietet – oder performativer Art: „To make the statement that ‚my [research] is conducted without values' is to hold at least one value" (Letherby et al., 2013, S. 59).

Befürworterinnen und Befürwortern von Wertfreiheit können diesem Problem auf zwei Weisen begegnen. Entweder sie konzedieren, dass $\mathrm{WFI}_{\mathrm{Norm}}$ eine homologische Regel ist, versuchen aber zu zeigen, dass $\mathrm{WFI}_{\mathrm{Norm}}$ dennoch konsistent ist; oder sie argumentieren, dass $\mathrm{WFI}_{\mathrm{Norm}}$ tatsächlich eine heterologische Regel ist, wodurch sich das Problem der Selbstanwendung gar nicht erst ergibt. Betrachten wir zunächst die zweite Option. Ein möglicher Ausgangspunkt hierfür könnte es sein, Wertfreiheit als *konstitutiven* Begriff zu verstehen (d.h. Wertfreiheit würde genuin wissenschaftliche Aktivitäten überhaupt erst ermöglichen). Ähnliches ist aus anderen Zusammenhängen bekannt. So können etwa Abstimmungsregeln nicht selbst qua Abstimmung generiert werden, weil hierfür bereits eine solche Regel vorausgesetzt werden muss. Abstimmungsregeln sind prinzipiell heterologisch – sie konstituieren ihren Geltungsbereich, ohne selbst in ihm zu liegen. Ähnliche Denkfiguren finden sich bei Carnap und Reichenbach. Carnap (1950) unterschied zwischen *internen* Fragen, die innerhalb eines semantischen Systems beantwortbar sind (z.B. „ist ‚fünf' eine Zahl?"), und *externen* Fragen, die sich auf das System als Ganzes richten (z.B. „sind Zahlen reale Entitäten"?). Für Carnap sind externe Fragen unbeantwortbar, weil die hierfür nötigen Regeln nur innerhalb eines Systems gelten und nicht für das System selbst. Carnaps semantische Systeme sind somit Systeme heterologischer Regeln. Bei Reichenbach ist das heterologische Regelverständnis die Lösung

dass es konsistente und inkonsistente homologische Regeln gibt. Anders als bei Grelling und Nelson handelt es sich bei den inkonsistenten Regeln jedoch nicht um Paradoxien, sondern um einfache Selbstwidersprüche (hier ist Poppers Formulierung, unbedingte Wertfreiheit sei „ein Paradox" (Popper, 1974, S. 115), begrifflich irreführend).

für das Problem, dass das Kriterium der Verifizierbarkeit (ein zentrales Konzept seiner Bedeutungstheorie) selbst nicht verifizierbar ist: „the question of meaning is not a matter of truth-character but […] a volitional decision" (1938/1961, S. 62). Für Reichenbach konstituiert die Verifikationsregel somit den Bereich wahrer und falscher Aussagen, ohne selbst wahr oder falsch zu sein.

Ist es nun denkbar, dass WFI_{Norm} in derselben Weise heterologisch ist, wie eine Abstimmungsregel oder die konstitutiven Regeln Carnaps und Reichenbachs? Ich denke, die Antwort hierauf lautet „nein". Denn diese Argumentation beruht offenbar auf der nicht-trivialen – und wahrscheinlich falschen – Annahme, dass wertbeladene Wissenschaft nicht nur *schlechte* Wissenschaft ist, sondern *gar keine*. Eben dies würde es nämlich bedeuten, wenn WFI_{Norm} seinen Anwendungsbereich zu allererst konstituieren würde. Das scheint jedoch nicht überzeugend, denn dann wäre der Anteil der Forschungsergebnisse, die tatsächlich als Wissenschaft gelten können, selbst aus Sicht der Befürworterinnen und Befürworter des Wertfreiheitsideals minimal. Insbesondere für die im letzten Kapitel beschriebenen Fälle *entschuldbarer* Wertbeladenheit sowie die erwähnten Beispiele aus der Wissenschaftsgeschichte scheint dieses Wissenschaftsverständnis unangemessen restriktiv.

Eine Erwiderung könnte darin bestehen, die Selbstanwendung von WFI_{Norm} schlicht definitorisch auszuschließen („WFI_{Norm} gelte für alle nicht-epistemischen Werturteile, jedoch nicht für sich selbst"). Anders als die definitorische Einschränkung auf kontextuelle und differenzielle Wertfreiheit (s. Kap. 5) scheint dies jedoch arbiträr (ähnlich wie „das Verbot diskriminierender Sprache gelte für alle Sprechakte außer für sich selbst"). Allerdings wirft diese Erwiderung eine interessante Frage auf: Ist WFI_{Norm} tatsächlich ein *nicht-epistemisches* Werturteil? Wenn dies nicht der Fall wäre, könnten Verteidigerinnen und Verteidiger von Wertfreiheit die erste Option wählen, nach der WFI_{Norm} zwar homologisch, aber dennoch konsistent ist. Das, was WFI_{Norm} verbietet, wäre somit etwas anderes als das, was mit WFI_{Norm} vollzogen wird. Diesen Weg wählt etwa Hugh Lacey:

> [I]t is a value that science be value-free. No irony or paradox is intended. The value desired to be reflected in sound theory choice is not moral, personal, social or aesthetic […]. It is cognitive value (Lacey, 1999, S. 55).

Wenn WFI_{Norm} jedoch ein epistemisches Werturteil ist, darf dieses nicht (oder nicht vorrangig) ethisch, sozial oder politisch motiviert sein. Denn gemäß WFI_{Diff} sollen epis-

temische Werturteile ja in ihrer Funktion, wissenschaftliche Erkenntnis zu befördern, hinreichend von nicht-epistemischen Werturteilen unterscheidbar sein (s. Kap. 5). Das heißt zwar nicht, dass Wertfreiheit darüber hinaus nicht auch weiteren (politischen, ethischen etc.) Zielen dienen dürfte; jedoch müssten die epistemischen Ziele alleine genommen hinreichend zur Begründung von WFI_{Norm} sein. Häufig diskutierte epistemische Motive sind etwa die Verhinderung von Wunschdenken und Verzerrungen, die Frage der Relevanz nicht-epistemischer Werturteile für die Wahrheitssuche oder das Problem der Begründbarkeit nicht-epistemischer Werturteile (z.B. Weber, 1917/1988; Douglas, 2000; Haack, 2003; Anderson, 2004; Wilholt, 2009; Schurz 2013). Als nicht-epistemische Motive werden oft die Abwehr von Expertenherrschaft, das Unterbinden unfairer Bevorteilung sowie die Anforderung diskutiert, dass Wissenschaft für Stakeholder mit unterschiedlichen Werten und Zielen gelten können soll (z.B. Betz, 2013; De Melo-Martin & Intemann, 2016; Bright, 2018; Holman & Wilholt, 2022).

Da ich einige dieser Motive an anderer Stelle diskutiere (s. Teil I und Teil III dieser Arbeit), werde ich die epistemischen und nicht-epistemischen Quellen des Wertfreiheitsideals hier nicht weiter ausführen. Wichtig ist jedoch, dass beide Arten von Motiven in der Literatur eine mindestens gleichwertige Rolle spielen; darüber hinaus hatten nicht-epistemische Begründungen einen wichtigen Anteil an der Entstehungsgeschichte des Wertfreiheitsideals (Proctor, 1991; Douglas, 2009, Kap. 3; Büter, 2012). Vor diesem Hintergrund scheint es nicht ratsam, WFI_{Norm} auf epistemische Motive zu reduzieren. Zumindest scheint ein solches Verständnis enger zu sein, als es angesichts der verfügbaren Argumente und der Geschichte des Wertfreiheitsideals angemessen wäre. Damit ist zwar nicht gesagt, dass dies im strengen Sinne ausgeschlossen wäre, jedoch würde der Gewinn an Konsistenz um den Preis eines gewissermaßen verarmten Wertfreiheitsbegriffs erreicht werden.

Es gibt allerdings einen weiteren Ansatz, der ohne diesen Nachteil auskommt. Möglicherweise ist WFI_{Norm} deswegen eine konsistente homologische Regel, weil die nicht-epistemischen Werturteile, die mit WFI_{Norm} vollzogen werden, *anderer Art* sind als die nicht-epistemischen Werturteile, die von WFI_{Norm} verboten werden. Dies ähnelt Laceys Überlegung, dass sich in der Formulierung „der Wert von Wertfreiheit" die Begriffe „Wert" und „Wertfreiheit" nicht auf dasselbe beziehen. Anders als bei Lacey geht es hier aber nicht auf den Unterschied zwischen epistemischen und nicht-epistemischen Werturteilen, sondern um den zwischen verschiedenen nicht-epistemischen Werturteilen.

Eine vielversprechende Strategie besteht darin, nicht-epistemische Werturteile *erster* und *zweiter Ordnung* zu unterscheiden. Urteile erster Ordnung sind diejenigen, die Personen und soziale Gruppen zu konkreten normativen Fragen vertreten, etwa zu Fragen der sozialen Gerechtigkeit oder der Ökologie, aber auch zu tagespolitischen Themen. Diese Werturteile prägen die sozialmoralischen Identitäten dieser Personen und Gruppen und grenzen sie so von anderen Identitäten ab. Nicht-epistemische Werturteile zweiter Ordnung betreffen hingegen die normative Ordnung, innerhalb derer die Werturteile erster Ordnung miteinander in Beziehung treten. Beispiele hierfür sind die klassischen Freiheitsrechte liberaler Demokratien. Diese ermöglichen es, innerhalb eines gewissen Rahmens unterschiedliche – und häufig konfligierende – Werte zu vertreten. Obwohl diese Rechte nicht wert*frei* sind, sind sie doch wert*neutral* in dem Sinne, dass sie unparteilich gegenüber diversen Konzeptionen des guten Lebens sind. Auf ähnliche Weise scheinen die WFI_{Norm} motivierenden Urteile, zumindest dem Anspruch nach, unparteilich gegenüber konkreten ethischen oder politischen Positionen zu sein. Augenfällig ist dies etwa für das Werturteil, „that collective goals are [to be] determined by democratically legitimized institutions, and not by a handful of experts" (Betz, 2013, S. 207) sowie für das Urteil, dass Wissenschaft für alle Stakeholder, d.h. unabhängig von partikularen Werten und Zielen gelten können soll (vgl. z.B. Weber, 1904/1988, S. 184; Lacey, 1999, S. 74).

Natürlich ist auch dieser Ansatz nicht völlig risikofrei. So müssten Verteidigerinnen und Verteidiger von Wertfreiheit neben der bereits verwendeten Unterscheidung (epistemisch vs. nicht-epistemisch) eine weitere potenziell angreifbare Unterscheidung (erste vs. zweite Ordnung) einführen. Ich lasse an dieser Stelle offen, ob diese Strategie letztlich erfolgreich wäre. Zumindest scheint sie plausibel genug, um das Selbstanwendungsproblem *prima facie* als unschädlich für WFI_{Norm} anzusehen. Wichtig ist dabei auch, dass die skizzierte Argumentation nicht behaupten muss, dass die Effekte des Wertfreiheitsideals für alle sozialen Gruppen gleich gut sein werden; zur Abwehr des Selbstwidersprüchlichkeitseinwands reicht es vielmehr aus, dass der nicht-epistemische Wert von Wertfreiheit systematisch von den nicht-epistemischen Werturteilen verschieden ist, die das Wertfreiheitsideal aus dem epistemischen Kernbereich ausschließen will. Da dies zunächst plausibel erscheint, kann WFI_{Norm} durchaus als homologische, aber konsistente Regel verstanden werden.

4. Die deskriptive These (WFI_{Desk})

4.1 DER TERMINUS „SIGNIFIKANT“

In diesem Kapitel diskutiere ich WFI_{Desk}, also den Möglichkeitsanspruch des Wertfreiheitsideals. Nach WFI_{Desk} ist es prinzipiell möglich, alle signifikanten nicht-epistemischen Werturteile aus dem epistemischen Kernbereich auszuschließen (wobei, wie im vorangegangenen Kapitel argumentiert, hiermit nicht-epistemische Werturteile erster Ordnung gemeint sind). Im Folgenden betrachte ich zwei Aspekte dieser These: die Einschränkung auf *signifikante* Werturteile sowie die Einschränkung auf die *prinzipielle* Möglichkeit von Wertfreiheit. Ich beginne mit dem Terminus „signifikant“.

Wie bereits erwähnt (s. Kap. 2.2), bezieht sich dieser Terminus auf den Unterschied zwischen Werturteilen, die einen maßgeblichen Einfluss auf die Wissenschaft haben, und Werturteilen, deren Einfluss geringfügig oder zumindest beherrschbar ist (etwa durch nachträgliche Korrekturen). Ein denkbares Kriterium hierfür ist die Frage, ob ein Werturteil die finalen Ergebnisse eines Forschungsprozesses tatsächlich verzerrt und ob diese Verzerrung umfangreich genug ist, um in dem gegebenen Kontext einen tatsächlichen Unterschied zu machen. Wenn dies der Fall ist, kann von signifikanter Wertbeladenheit gesprochen werden. Da sich WFI_{Desk} lediglich auf diese Art von Wertbeladenheit bezieht, handelt es sich – ähnlich dem Terminus „prinzipiell“ sowie dem Begriff der „partiellen“ (i.e. kontextuell-differenziellen) Wertfreiheit – um eine definitorische Einschränkung. Allerdings wird diese Einschränkung in der Literatur nur selten diskutiert. Ich halte es dennoch für sinnvoll, sie explizit zu erwähnen, da sie ein naheliegendes Missverständnis ausräumt: Eine Kritik an WFI_{Desk} könnte versucht sein, Maßnahmen zur Realisierung von Wertfreiheit als unzureichend zurückzuweisen, sobald diese nicht *uneingeschränkt* erfolgreich sind. Jeder noch so unwesentliche Werteinfluss wäre dann schlagend gegen WFI_{Desk}. Ein solcher Maßstab scheint jedoch unangemessen strikt. Diese Intuition steht etwa hinter einer Kritik von Katie Steele an Helen Longino:

> [T]here is a large difference between social-political background subtly affecting scientists judgments of simplicity that in turn affect their beliefs about hypotheses and the type of value judgments that […] explicitly take account of real-world consequences (Steele, 2012, S. 897).

Steele weist hier auf den Unterschied zwischen Werturteilen mit geringem oder vagem Effekt und solchen Werturteilen hin, die ein Forschungsergebnis unmittelbar und er-

heblich beeinflussen. Zwar argumentiert Steele im Weiteren gegen das Wertfreiheitideal, allerdings mit anderen Gründen und beschränkt auf wissenschaftliche Politikberatung. Ein verwandtes Argument findet sich in einem Beitrag von Wendy Parker. Obwohl Parker das gegenteilige Argumentationsziel verfolgt – die Verteidigung des Wertfreiheitsideals – stimmt sie mit Steele darin überein, dass das Vorhandensein insignifikanter Wertbeladenheit nicht zur Widerlegung von WFI_{Desk} ausreicht. Worauf es ankommt, ist der Effekt der Werturteile auf die fraglichen Forschungsergebnisse. So argumentiert Parker, dass Unsicherheiten in den Klimawissenschaften nicht exakt quantifiziert, sondern in gröberen, qualitativen Begriffen kommuniziert werden sollten. Ziel ist es dabei, den Einfluss von Werturteilen, die in die Entwicklung von Klimamodellen eingegangen sind, nachträglich unschädlich zu machen:

> [E]ven if social values sometimes do come into play in the model development process […], the influence of those values on estimates of uncertainty will be reduced when coarser estimates are given. The influence will be reduced insofar as choices in model development will less often *make a difference* to the uncertainty estimates produced (Parker, 2014, S. 28, Hervorh. i. Orig.).

Mir geht es an dieser Stelle nicht darum, ob die von Parker vorgeschlagene Strategie tatsächlich funktioniert. Wichtig ist hier lediglich, dass mit dem Nachweis nicht-epistemischer Werteinflüsse auf den epistemischen Kernbereich noch nicht viel für eine Kritik an WFI_{Desk} gewonnen ist. Hinzukommen muss der Nachweis relevanter Auswirkungen auf die entsprechenden Forschungsergebnisse. Wenn es etwa plausibel erwartbar ist, dass zwei konkurrierende, aber empirisch ebenbürtige Modelle in einem gegebenen Kontext zu ähnlichen Ergebnissen führen werden, dann scheint die Wahl zwischen diesen Modellen auch dann keine signifikante Wertbeladenheit zu erzeugen, wenn sie aufgrund nicht-epistemischer Werturteile getroffen wurde. Ähnlich wäre es, wenn Werturteile zwar zunächst signifikant sind, im Nachhinein jedoch erkannt und auf ein insignifikantes Maß vermindert werden können. Eine Kritik an WFI_{Desk} benötigt daher ein Argument, das den ergebnisändernden Charakter nicht-epistemischer Werturteile nachweist – nicht eines „[that] exaggerates the influence of social values" (Parker, 2014, S. 27).

4.2 DER TERMINUS „IM PRINZIP"

Die zweite Einschränkung von WFI_{Desk} wird durch den Ausdruck „im Prinzip" oder „prinzipiell" angezeigt. Diese Einschränkung wird in der Literatur häufig erwähnt:

> [The value-free ideal claims that] [s]cientists can – at least in principle – gather evidence and assess/accept theories without making contextual value judgments (Reiss & Sprenger 2020).
> [T]he charge [against value-freedom] is essentially about the *theoretical impossibility* of value-neutral results. It is not only about what doesn't happen *in practice* in science, it is about what cannot happen even *in principle*, in ideal cases of properly conducted research" (Ruphy, 2006, S. 191–192, Hervorh. i. Orig.).
> [I]n principle […] theories can be accepted […] solely in the light of the empirical data, other accepted theories and the play of the cognitive values (Lacey, 1999, S. 251).

Der Sinn dieser Einschränkung besteht darin, WFI_{Desk} gegen Angriffe abzusichern, die von der empirisch vorfindlichen Wertbeladenheit wissenschaftlicher Forschung auf die strikte Unmöglichkeit von Wertfreiheit schließen. Derartige Angriffe beziehen sich auf kontingente und daher änder- oder kompensierbare Hindernisse für Wertfreiheit. Hierzu zählen etwa das historisch gewachsene wissenschaftliche Institutionengefüge, die Verteilung von Ressourcen und Interessen in der Wissenschaft, kulturelle Gegebenheiten und Ähnliches. Auch andere kontingente Umstände, die zwar nicht änderbar sind, aber kompensiert werden können, fallen in diese Kategorie. Einschlägig sind hier etwa die allgemeine Fehlbarkeit des Menschen oder der Tatbestand, dass Wissenschaftlerinnen und Wissenschaftler nicht nur epistemische, sondern auch moralische und politische Subjekte sind. Derartige Hemmnisse sind fraglos vorhanden, und vielfach führen sie *de facto* zu Wertbeladenheit. Wenn sie also für eine Kritik von WFI_{Desk} uninteressant sind, so nicht deshalb, weil sie für Vertreterinnen und Vertreter des Wertfreiheitsideals unproblematisch wären. Im Gegenteil: Aus ihrer Sicht sind sie mindestens ein Anlass zur Sorge, wenn nicht gar zur Zurückweisung der jeweiligen Forschungsergebnisse (wenn es sich um Fälle nicht entschuldbarer Wertbeladenheit handelt).

Wieso sind praktische Realisierungsprobleme dennoch kompatibel mit WFI_{Desk}? Die Antwort hierauf ist zweiteilig. Der erste Grund besteht darin, dass, solange Wertbeladenheit kontingenter Natur ist, Kritiken an WFI_{Desk} stets mit der Forderung begegnet werden können „dass wir uns mehr anstrengen sollten" (Koertge, 2013, S. 247). Denn die prinzipielle Möglichkeit von Wertfreiheit bedeutet ja nichts anders, als dass die Überwindung existierender Wertbeladenheit nicht grundsätzlich ausgeschlossen ist, etwa indem mehr Ressourcen oder neuartige Maßnahmen mobilisiert werden. Diese Maßnahmen können durchaus radikal sein. Ein Beispiel hierfür ist James Browns (2013) Kritik der medizinischen Forschungslandschaft. Diese Forschung, so Brown, ist durch

die nicht-epistemischen Interessen der pharmazeutischen Industrie stark verzerrt. Anstatt jedoch den Anspruch auf unverzerrte Medizin aufzugeben, fordert Brown „nicht bloß eine strengere Anwendung der existierenden Methoden guter Wissenschaft, sondern eine durch politisches Handeln erreichte Neuorganisation wissenschaftlicher Forschung“ (ebd., S. 338). Dies beinhaltet die Abschaffung medizinischer Patente und einen massiven Ausbau öffentlicher Finanzierung. Die Überlegung Browns, der nicht explizit von Wertfreiheit, sondern von „objektive[r] Wissenschaft“ (ebd., S. 362) spricht, ist insofern analog, als die zweifellos vorhandenen Hemmnisse bei der Realisierung seines Vorschlags nicht gegen den Vorschlag selbst sprechen. Eben hierauf weist der Terminus „im Prinzip“ hin: dass selbst schwerwiegende Hindernisse nicht die Unmöglichkeit von Wertfreiheit implizieren.

Es ist jedoch noch ein zweites Argument für die Immunität von WFI_{Desk} gegenüber kontingenter Wertbeladenheit denkbar. In einer bestimmten Hinsicht muss Wertfreiheit selbst dann, wenn sogar radikale Schritte zu ihrer Realisierung aussichtslos wären, nicht strikt unmöglich sein. Denn es ließe sich argumentieren, dass dieses Scheitern eher etwas über *uns* aussagt als über die wissenschaftliche Erkenntnis *als solche*. Ein derartiges Argument würde auf dem Unterschied zwischen den Methoden der Erkenntnis und unserer Fähigkeit beruhen, diese adäquat anzuwenden. Sollten wir über letztere nicht oder nur eingeschränkt verfügen, so das Argument, würde dies lediglich implizieren, dass *wir* nicht wertfrei forschen können – die ideale Wissenschaftlerin oder der ideale Wissenschaftler könnte dies durchaus. Eine solche Überlegung findet sich in Richard Rudners (1953) berühmter Darstellung des Wertfreiheitsideals:

> The perfect scientist – the scientist *qua* scientist does not allow this kind of value judgment to influence his work. However much he may find doing so unavoidable *qua* father, *qua* lover, *qua* member of society, *qua* grouch, when he does so he is not behaving *qua* scientist (Rudner, 1953, S. 2).

Diese Denkfigur scheint die Möglichkeit zu eröffnen, dass selbst *unvermeidliche* Wertbeladenheit WFI_{Desk} nicht widerlegt – zumindest dann nicht, wenn die Unvermeidlichkeit nicht in die idealisierte Sphäre „perfekter“ Wissenschaft hineinragt (ich diskutiere diese Frage in Teil III dieser Arbeit). Die Überlegung hat jedoch einige Schwächen. Wenn prinzipielle Möglichkeit bereits dann gegeben ist, wenn Wertfreiheit lediglich logisch möglich ist, dann ist unklar, ob hieraus überhaupt etwas für den realen Forschungsbetrieb folgt. Denn zum einen scheint das bereits diskutierte Sollen-impliziert-Können dann nur auf ideale, nicht aber auf tatsächliche Wissenschaftlerinnen und Wissenschaft-

ler anwendbar zu sein; zum anderen scheinen die Motive des Wertfreiheitsideals (Verhinderung von Wunschdenken, Abwehr von Expertenherrschaft etc.) die Geltung des Ideals für tatsächliche Forschung dann nicht mehr begründen zu können. Aus Sicht von Vertreterinnen und Vertretern des Wertfreiheitsideals kann es daher aussichtsreicher sein, den Terminus „im Prinzip" eher im ersten diskutierten Sinn, d.h. als Möglichkeit von Wertfreiheit bei Zuhilfenahme umfangreicher Ressourcen und radikaler Maßnahmen zu verstehen.

Dennoch denke ich, dass die Denkfigur des „perfect scientist" hilfreich für die Diskussion ist. Denn selbst wenn die Unterscheidung zwischen möglicherweise perfekten, aber nicht existierenden Wissenschaft und der unperfekten Forschungsrealität unbefriedigend erscheint, so macht sie doch auf einen wichtigen Tatbestand aufmerksam: Die Wertfreiheit des „scientist *qua* scientist" oder der „scientific method as such" (Rudner, 1953, S. 2) scheint auf einer fundamentaleren Ebene zu liegen als die tatsächlichen Bedingungen realer Forschung. Kann daher gezeigt werden, dass nicht erst die Methodenanwendung, sondern bereits die Methoden *selbst* wertbeladen sind, so lässt sich hieraus ein besonders starkes Argument gegen WFI_{Desk} herleiten: „idealizations […] strengthen the conclusion because it is more surprising in the ideal setting that scientists must make value judgments" (Steele, 2012, S. 895). Die Denkfigur der perfekten Wissenschaft erlaubt es daher, die begrifflichen Stärken und Schwächen wertfreiheitskritischer Argumente zu testen (s. Teil III dieser Arbeit; dort zeige ich auch, wie die Figur des „perfect scientist" so operationalisiert werden kann, dass sie auch für reale Wissenschaftlerinnen und Wissenschaftler informativ ist).

5. Geltungsbereich des Wertfreiheitsideals: Forschungskontexte und Werturteilsarten

5.1 DIE KONTEXTUALITÄTSTHESE (WFI_{KONT})

Während WFI_{Norm} und WFI_{Desk} die Geltungsansprüche des Wertfreiheitsideals repräsentieren, beziehen sich die verbleibenden beiden Teilthesen auf seinen Geltungsbereich. Im folgenden Kapitel diskutiere ich zunächst WFI_{Kont}, die Kontextualitätsthese, um dann im zweiten Schritt die Differenzialitätsthese WFI_{Diff} zu betrachten. Beide Teilthesen haben in der Literatur vielfältige Kritik auf sich gezogen (z.B. Longino, 1996;

Machamer & Douglas, 1999; Putnam, 2002; Dupré, 2007; Bueter, 2015; De Melo-Martin & Intemann, 2016). Thema dieses Kapitels ist jedoch weniger, ob WFI_{Kont} und WFI_{Diff} gültig sind, sondern was sie bedeuten und wie sie motiviert sind. Ich lasse daher die einschlägigen Kritiken außer Acht und konzentriere mich auf den Inhalt und Hintergrund der Teilthesen.

Wie ich bereits betont habe, ist das Wertfreiheitsideal auf einen bestimmten Bereich oder Kontext beschränkt: den epistemischen Kernbereich von Wissenschaft. WFI_{Kont} behauptet, dass dieser Bereich hinreichend von anderen Wissenschaftsbereichen abgrenzbar ist. In variierender Terminologie wird diese Unterscheidung häufig in der Literatur erwähnt:

> Ein bestimmter Bereich der Wissenschaften, nämlich ihr *Begründungszusammenhang*, soll frei sein von fundamentalen wissenschaftsexternen Wertannahmen (Schurz, 2013, S. 313, Hervorh. i. Orig.).

Dem liegt die Vorstellung zugrunde, dass nicht alles, was in Forschungsprozessen geschieht, erkenntniskonstitutiv ist – und dass folglich nicht alles, was Wissenschaftlerinnen und Wissenschaftler tun, durch das Wertfreiheitsideal reguliert werden muss. In diesem Sinne argumentiert etwa Hans Reichenbach, einer der Begründer der Bereichsunterscheidung:

> [T]here are certain elements of knowledge, however, which are not governed by the idea of truth, but which are due to volitional resolutions, and though highly influencing the makeup of the whole system of knowledge, [these elements] do not touch its truth-character (Reichenbach, 1938/1961, S. 9).

Um diesem Unterschied gerecht zu werden, muss aus Sicht des Wertfreiheitsideals in verschiedene Wissenschaftsbereiche oder -kontexte[14] unterschieden werden: zum einen in einen Kernbereich, der „at the heart of the research process" (Douglas, 2009, S. 45) liegt und der somit den Wahrheitsanspruch von Wissenschaft begründet. Typische Aktivitäten in diesem Bereich sind etwa die Sammlung und Auswertung von Daten, die Wahl von Methoden und Modellen sowie die Akzeptanz oder Zurückweisung von Hypothesen und Theorien. Dieser Bereich wird auch als „Begründungs-" oder „Rechtfertigungs-

14 Hier besteht eine Verwechslungsgefahr: In der Wertfreiheitsdebatte wird manchmal von „contextual values" oder „contextual factors" gesprochen. Nach der von mir verwendeten Terminologie handelt es sich hierbei um nicht-epistemische Werte. Diese sind jedoch Gegenstand der *Differenzialitätsthese* WFI_{Diff} und nicht, wie die Bezeichnung „contextual" vermuten lässt, der *Kontextualitätsthese* WFI_{Kont}.

zusammenhang", „interner Bereich" und „interne Phase" bezeichnet. Zum anderen existieren, so die Idee, „externe" Bereiche, in denen wissenschaftliches Wissen nicht *qua* Wissen konstituiert, sondern lediglich motiviert, angewandt oder vermittelt wird. Diese werden auch als „Entdeckungs-", „Anwendungs-" oder „Kommunikationszusammenhänge" bezeichnet und umfassen die Auswahl von Forschungsfragen, die Verwendung von Wissenschaft für praktische Zwecke und die Dissemination von Forschungsergebnissen. Auch ethische Standards, wie sie etwa in klinischen Studien angewandt werden, zählen in den externen Bereich (Douglas, 2009).

Wie ist diese Trennung in verschiedene Wissenschaftsbereiche nun zu verstehen? Paul Hoyningen-Huene (2006) identifiziert fünf verschiedene Interpretationen, wobei die zentrale Frage darin besteht, ob die Unterscheidung auf der Objektebene tatsächlich ablaufender Forschung oder der Metaebene wissenschaftsphilosophischer Analyse operieren[15]. Für die vorliegende Diskussion möchte ich dies in abgewandelter Form übernehmen und zwischen einer *logischen*, *rekonstruktiven*, *diachronen* und *synchronen* Lesart der Bereichsunterscheidung differenzieren. Die logische Lesart liegt auf der Metaebene und interpretiert die Wissenschaftsbereiche als begrifflich distinkte Typen von Fragen: einerseits normative Fragen der epistemischen Gerechtfertigtheit eines Forschungsergebnisses (Kernbereich), andererseits deskriptive Fragen seiner empirischen Bedingtheit, etwa hinsichtlich der Entstehungsgeschichte eines Forschungsergebnisses (externer Bereich). Da sich diese Lesart jedoch nur *ex post*, d.h. aus der Perspektive der wissenschaftsphilosophischen Analyse, nicht aber *in situ*, also aus der Perspektive des Forschungsprozesses selbst anwenden lässt, ist die logische Lesart nicht zur Spezifikation von $\mathrm{WFI}_{\mathrm{Kont}}$ geeignet. Vielmehr muss die Unterscheidung auf der Objektebene liegen, wenn sie reale Forschungsentscheidungen informieren können soll.

Dasselbe gilt für die rekonstruktive Lesart. Diese geht auf Hans Reichenbach (1938/1961) zurück und versteht den Kernbereich (den „Rechtfertigungskontext" in Reichenbachs Terminologie) als Ergebnis einer philosophischen Analyse, die alles aus einem Forschungsergebnis eliminiert, was nicht durch rein epistemische Prinzipien legitimiert ist: „we replace actual thinking by such operations as are justifiable, that is, as can be demonstrated as valid" (ebd., S. 7). Der Kernbereich fungiert hier ebenfalls nicht als

15 Die Diskussion, auf die sich Hoyningen-Huene hier bezieht, liegt allerdings quer zu meiner – während es mir um eine Definition des Wertfreiheitsideals geht, geht es Hoyningen-Huene insbesondere darum, Aufgaben und Grenzen der Wissenschaftsphilosophie festzulegen. Vgl. auch die Einleitung von Jutta Schickore und Friedrich Steinle in demselben Band: „For several decades, the context distinction dictated what philosophy of science should be" (Schickore & Steinle, 2006, S. vii).

Teil tatsächlicher Forschungsprozesse, sondern als ex-post-Rationalisierung: „we do not maintain anything about the question of how [science] is performed – what we maintain is nothing but a relation of a theory to facts, independent of the man who found the theory" (S. 382). Aus Sicht des Wertfreiheitsideals ist hiermit jedoch nichts gewonnen, denn dessen Thema ist nicht, zu welchen Ergebnissen Wissenschaftlerinnen und Wissenschaftler unter kontrafaktischen Bedingungen gelangt *wären*, sondern ihre *tatsächlichen* Pflichten und Möglichkeiten. Somit ist auch die rekonstruktive Lesart der Bereichsunterscheidung nicht zur Spezifikation von $\mathrm{WFI}_{\mathrm{Desk}}$ geeignet.

Im Gegensatz zu den ersten beiden Interpretationen operieren die diachrone und die synchrone Lesart auf der Objektebene. Der epistemische Kernbereich umschreibt hierbei die bereits erwähnten Aktivitäten (z.B. Datensammlung, Modellwahl, Hypothesenprüfung), wobei der Hypothesen- und Theorienprüfung ein besonderer Stellenwert zukommt. Die diachrone Lesart besteht darin, diese Aktivitäten in eine zeitliche Ordnung zu bringen, so dass eine wertbeladene „discovery period" (Mowry, 1985, S. 79) von einer wertfreien „justification period" (ebd.) abgelöst wird. Erstere beinhaltet die Festlegung von Forschungsfragen, letztere die Hypothesenbewertung, also „the subsequent testing of and giving good reasons for [a] theory" (ebd.). Hierauf folgen weitere wertbeladene Phasen wie die Anwendung und Vermittlung des Forschungsergebnisses. Bryan Mowry bezeichnet diese Lesart als „standard formulation" (ebd.). In ähnlicher Weise verortet auch die synchrone Lesart die Bereichsunterscheidung auf der Objektebene. Hierbei lässt sie jedoch Iterationen zu, so dass sich interne und externe Aktivitäten abwechseln können. Der Forschungsprozess erscheint dann weniger als Abfolge von Phasen, sondern als Nebeneinander qualitativ verschiedener Handlungen. Ob eine Forschungshandlung zum epistemischen Kernbereich gehört, bestimmt sich nach der diachronen Lesart also nicht nach ihrem *Zeitpunkt*, sondern ihrem *Zweck*.

Da der epistemische Kernbereich sowohl nach der diachronen als auch nach der synchronen Lesart Situationen beschreibt, in denen sich Wissenschaftlerinnen und Wissenschaftler tatsächlich befinden, sind beide zur Spezifikation von $\mathrm{WFI}_{\mathrm{Kont}}$ geeignet. Unabhängig davon, welche der beiden Interpretationen gewählt wird, gilt: „The ideal of value-freedom is supposed to apply when scientists are deciding whether or not to accept some hypothesis (or theory) on the basis of evidence" (Kitcher, 2011, S. 32). Da Wissenschaft viele Aktivitäten beinhaltet, die nicht unter diese Beschreibung fallen, eröffnen beide Lesarten vielfältige Möglichkeiten für nicht-epistemische Werturteile, etwa ethischer Art. So vertritt etwa Noretta Koertge (2013) ein „Erklärung-plus-Ethik-Modell", das „vereinbar mit der Forderung nach einem hohen Grad an sozialer Verantwor-

tung“ ist (ebd., S. 239). Dennoch, so Koertge, sollen ethische Urteile keine Rolle in der Hypothesenprüfung spielen. So beschränken etwa forschungsethische Entscheidungen „nur die Methoden, mit denen Theorien überprüft werden können“ (ebd.), ohne dass deswegen ethische Werte in das wissenschaftliche „Buch der Tugenden“ (ebd.) aufgenommen werden müssen[16]. Aus Sicht des Wertfreiheitsideals ist es jedoch weniger wichtig, ob der epistemische Kernbereich auf diachrone oder synchrone Weise von den externen Bereichen getrennt ist; zentral ist vielmehr, dass diese Trennung auf der Objektebene tatsächlicher Forschung durchführbar ist. Ex-post-Perspektiven, wie sie etwa für Reichenbachs Verständnis der Bereichsunterscheidung kennzeichnend sind, sind hierfür nicht geeignet.

5.2 DIE DIFFERENZIALITÄTSTHESE (WFI_{DIFF})

Im Gegensatz zu WFI_{Diff}, die das *Wann* (diachrone Lesart) oder *Wozu* (synchrone Lesart) von Werturteilen thematisiert, richtet sich die Differenzialitätsthese WFI_{Diff} auf ihr *Was*, d.h. auf den Gehalt dieser Urteile. Hierfür differenziert WFI_{Diff} *epistemische* und *nicht-epistemische* Werturteile und behauptet, dass diese Typen von Urteilen hinreichend voneinander unterscheidbar sind. In WFI_{Diff} drückt sich somit, wie in WFI_{Kont}, eine Strategie der Einschränkung durch Abgrenzung aus.

Epistemische Werteurteile beziehen sich auf die Eigenschaften von Gegenständen, mit denen Wissenschaftlerinnen und Wissenschaftler im epistemischen Kernbereich umgehen. Hierzu zählen insbesondere Hypothesen und Theorien, aber auch Modelle, Methoden, Instrumente und ähnliches. Im Zentrum stehen dabei diejenigen Eigenschaften, die es wahrscheinlich erscheinen lassen, dass die fraglichen Gegenstände entweder selbst wissenschaftliches Wissen instanziieren (Theorien, Hypothesen) oder dabei helfen, wissenschaftliches Wissen zu generieren (Modelle, Methoden etc.). Je mehr dies der Fall ist, umso größer ist der epistemische Wert dieser Gegenstände: „An epistemic value is one we have reason to believe will, if pursued, help toward the attainment of [scientific] knowledge” (McMullin, 1982, S. 18). Da das wissenschaftliche Unternehmen maßgeblich dadurch motiviert ist, eben dieses Wissen zu erlangen, *dürfen* Wissenschaftlerinnen

16 Im Gegensatz zu anderen Vertreterinnen und Vertretern des Wertfreiheitsideals lässt Koertge nicht-epistemische Werturteile nur bezüglich ethischer Methodenbeschränkungen und der Anwendung von Forschungsergebnissen zu. Die Auswahl von Forschungsproblemen – traditionell ebenfalls Teil des externen Bereichs (s. etwa Weber, 1904/1988; Weber, 1917/1988) – beschränkt Koertge hingegen auf Fragen, „die intellektuell interessant sind und uns helfen, die Welt, in der wir leben, zu verstehen“ (Koertge, 2013, S. 243). Wie es scheint, soll wissenschaftliches Agenda-Setting für Koertge also allein durch epistemische Werturteile bestimmt werden.

und Wissenschaftler diese Werturteile nicht nur fällen, sie *sollen* es sogar. Folglich beschränkt sich das Wertfreiheitsideal auf nicht-epistemische Werturteile, d.h. auf solche, die dem Ziel der Wissensgenerierung nicht dienen. Aus Sicht des Wertfreiheitsideals ist der genaue Inhalt dieser Urteile jedoch irrelevant – was zählt, ist die fehlende Ausrichtung auf das „ultimate goal of research, which is true (or at least reliable) knowledge" (Douglas, 2009, S. 93).

Der locus classicus dieser Diskussion ist ein Aufsatz von Thomas S. Kuhn (1977). Kuhn nennt fünf epistemische Werte: Übereinstimmung von Vorhersage und Beobachtung (*accuracy*), Übereinstimmung einer Theorie mit sich selbst und anderen akzeptierten Theorien (*consistency*), Umfang des Anwendungsbereichs (*scope*), Menge der benötigten Annahmen und praktische Handhabbarkeit (*simplicity*) sowie Menge und Qualität der ermöglichten Anschlussforschung (*fruitfulness*). Diese Aufzählung ist, wie Kuhn selbst betont, nicht erschöpfend; weitere Kandidaten wären etwa Testbarkeit, Formalisierbarkeit oder Reproduzierbarkeit. Kuhns Wertekanon ist jedoch nicht nur unabgeschlossen, er ermöglicht auch unterschiedliche Interpretationen und Gewichtungen. So lässt etwa Kuhns *accuracy* offen, welche Abweichung zwischen Vorhersage und Beobachtung gerade noch akzeptabel ist und wie Konflikte zwischen diesem und anderen Werten – etwa Kuhns *scope* – aufzulösen sind. Epistemische Werte leiten daher Aktivitäten im epistemischen Kernbereich an, erzwingen aber häufig keine Entscheidungen: „two men fully committed to the same list of criteria for choice may nevertheless reach different conclusions" (Kuhn, 1977, S. 324).

Neben der Vagheit dieses Wertekanons ist ein weiterer Punkt augenfällig: Obwohl alle genannten Werteurteile durch das Ziel der Wissensgenerierung motiviert sind, sind nur zwei von ihnen, nämlich die Übereinstimmung einer Theorie mit sich selbst sowie mit den Beobachtungen, direkt mit dem Wissensbegriff verknüpft. Diese Werte werden meist als interne Konsistenz und, in Abweichung von Kuhns Terminologie, empirische Adäquatheit bezeichnet. Zwar lassen auch interne Konsistenz und empirische Adäquatheit gewisse Auslegungsspielräume zu; dennoch muss jedes Forschungsergebnis, das als Wissen gelten können soll, beide Eigenschaften in gewissem Grade aufweisen. Stéphanie Ruphy (2006) bezeichnet interne Konsistenz und empirische Adäquatheit daher als universell gültige wissenschaftliche Werte. Ähnlich betont Heather Douglas (2009): „It is because we care about having reliable empirical knowledge that scientific theories must be internally consistent and predictively competent" (ebd., S. 95)[17]. Ich möchte diese

17 Douglas (2009, S. 92) bezeichnet sie daher im Anschluss an Larry Laudan als „epistemische Kriterien" anstatt „epistemische Werte". Der Einfachheit halber verwende ich dennoch die Bezeichnung „Werte" bzw. „Werturteile".

beiden Eigenschaften daher als *primäre* epistemische Werte bezeichnen und sie den *sekundären* gegenüberstellen.

Sekundäre epistemische Werte sind, anders als die primären, keine notwendige Bedingung für Wissen: Eine Theorie kann gültig sein, obwohl sie nicht umfassend, einfach oder fruchtbar ist. Dass diese Werte dennoch zu den epistemischen gezählt werden, liegt an ihrer Nützlichkeit für die Erkenntnispraxis. So schonen etwa einfache Theorien kognitive Ressourcen, während fruchtbare Theorien die Menge der adressierbaren Forschungsfragen erhöhen. Obwohl diese Eigenschaften keine „requirements of truth" (Steel, 2010, S. 18) sind, fördern sie auf indirektem Wege „the attainment of truth" (ebd.). Sekundäre epistemische Werte haben somit den Charakter der Zusätzlichkeit[18]. Dennoch kann ihnen eine wichtige Funktion zukommen, etwa bei der Wahl zwischen gleichermaßen empirisch adäquaten und intern konsistenten Alternativen. Kuhn (1977; 1957/1980) zeigt dies anhand der Kopernikanischen Revolution. Ende des sechzehnten Jahrhunderts stimmten sowohl das Kopernikanische, als auch das Ptolemäische und das System Tycho Brahes teilweise mit den verfügbaren astronomischen Daten überein. In derartigen Situationen können sekundäre epistemische Werte wie Einfachheit (pro Kopernikus) und externe Konsistenz (pro Ptolemäus) die weitere Erforschung der Alternativen motivieren. Da diese Werte unterschiedlich gewichtet werden können, können mehrere Theorien gleichzeitig verfolgt und so das Risiko von Fehlentscheidungen auf die wissenschaftliche Gemeinschaft verteilt werden. Dies kann äußerst produktiv sein:

> What from one viewpoint may seem the looseness and imperfection of choice criteria […] may, when the same criteria are seen as values, appear an indispensable means of spreading the risk which the introduction or support of novelty always entails (Kuhn, 1977, S. 332).

Wie bereits erwähnt, ist die Trennung in epistemische und nicht-epistemische Werturteile vielfach kritisiert wurden (z.B. Longino, 1996; Machamer & Douglas, 1999). Auch ist keineswegs klar, ob klassische epistemische Werte wie Einfachheit oder Konsistenz mit umgebenden Theorien in der von Kuhn vorgeschlagenen Weise aufrechterhalten wer-

18 Manche Autorinnen und Autoren verknüpfen sekundäre epistemische Werte direkter mit dem Wahrheitsbegriff. So hält etwa Ruphy (2006) *theoretische Eleganz* unter bestimmten Umständen für epistemisch wertvoll: „when a theory has turned out to be empirically successful, then its aesthetic properties are granted truth-conducing value" (ebd., S. 211). Dennoch sind derartige Werte auch in diesem Ansatz sekundär, denn sie unterliegen einer Einzelfallprüfung, die ihrerseits von primären epistemischen Werten wie empirischer Adäquatheit abhängt.

den können. So argumentiert etwa Helen Longino (2008), dass im Gegenteil eher ontologische Heterogenität und Neuheit erstrebenswerte Eigenschaften wissenschaftlicher Theorien seien. Da es mir an dieser Stelle jedoch mehr um den Inhalt von WFI_{Diff} als darum geht, ob diese Teilthese aufrechterhalten werden kann, gehe ich hier nicht weiter auf diese Kritiken ein. Relevant ist im vorliegenden Zusammenhang nur, dass sowohl primäre als auch sekundäre epistemische Werturteile kompatibel mit dem Wertfreiheitsideal sind. Weiterhin eröffnen diese Werturteile Interpretations- und Gewichtungsmöglichkeiten. Vertreterinnen und Vertreter des Wertfreiheitsideals sind daher nicht auf ein deterministisches Wissenschaftsverständnis festgelegt, nach dem epistemische Werte Forschungsentscheidungen *erzwingen*.

6. Schlussfolgerung

Das Wertfreiheitsideal ist eine komplexe, auf mehreren Teilthesen, Einschränkungen und Unterscheidungen beruhende philosophische Behauptung. Mit dem vorliegenden Aufsatz habe ich eine Möglichkeit vorgestellt, wie sich diese Komplexität strukturieren und in seinen Einzelaspekten diskutieren lässt. Ziel war es, auf der Grundlage eines konsistenten Begriffsverständnisses eine zumindest *prima facie* plausible Position zu erarbeiten und in der Literatur zu verorten. Dabei zeigt sich, dass das Wertfreiheitsideal auf einer Strategie der *Einschränkung durch Abgrenzung* basiert: zum einen auf der bekannten Abgrenzung verschiedener Wissenschaftsbereiche und Werturteilsarten; zum anderen auf den meist eher *en passant* erwähnten Unterscheidungen zwischen kontingenter und prinzipieller sowie zwischen signifikanter und insignifikanter Wertbeladenheit. Hinzu kommen weitere, in der Literatur bislang nicht diskutierte Unterscheidungen, etwa zwischen homologischen und heterologischen Regelverständnissen oder zwischen nichtepistemischen Werturteilen erster und zweiter Ordnung. Diese Abgrenzungen zielen auf die Abwehr naheliegender Kritiken ab, die sich etwa an der Wertbeladenheit der Problemwahl, der Notwendigkeit epistemischer Werturteile oder an dem Umstand entzünden, dass Wissenschaft häufig *de facto* wertbeladen ist.

Der vorliegende Aufsatz hat jedoch zwei wichtige Ausnahmen von dieser Strategie aufgezeigt. Erstens sollte Wertfreiheit nicht, wie es mitunter in der Literatur angedeutet wird, als *bloßes Ideal*, also als lobenswertes Ziel ohne tatsächlichen Realisierungsanspruch verstanden werden. Dies würde dem Verpflichtungscharakter des Wertfreiheitsideals widersprechen und nicht mit dem Grundsatz übereinstimmen, dass Sollensan-

sprüche immer auch Möglichkeitsannahmen implizieren (Sollen-impliziert-Können). Die Möglichkeitsthese wird auch nicht durch die de facto bestehende Wertbeladenheit von Wissenschaft widerlegt – zumindest dann nicht, wenn diese Wertbeladenheit nicht *prinzipieller* Natur ist. Damit zusammen hängt, dass, zweitens, das schwächere Verständnis des Wertfreiheitsideals als *approximatives* Ideal nicht unbedingt einer stärkeren Lesart überlegen ist, nach der die Erfüllungsbedingungen des Ideals das tatsächliche Erreichen von Wertfreiheit beinhalten. Diese stärkere Lesart lässt sich mit einem Konzept der Entschuldigungsgründe kombinieren, so dass nicht erreichte Wertfreiheit nicht zur grundsätzlichen Ablehnung eines Forschungsergebnisses führen muss. Obwohl ich mich an dieser Stelle nicht darauf festlege, welche der beiden Lesarten letztendlich erfolgreicher ist – ich selbst verwende in Teil III eine Formulierung, die eher der schwächeren Lesart ähnelt (allerdings ohne Konsequenzen für das dort vorgestellte Argument) – wäre es falsch, das Wertfreiheitsideal von vornherein mit der schwächeren Lesart zu identifizieren.

Offenbar laufen jedoch beide Vorschläge, die Integration des Möglichkeitsanspruchs und die stärkere Lesart des Sollensanspruchs, auf eine Erweiterung des Wertfreiheitsideals im Vergleich zu einer minimalistischen Interpretation hinaus. Obwohl dies von der Strategie der Einschränkung durch Abgrenzung abweicht, hat dieses umfassendere Verständnis Vorteile. Umgekehrt eröffnet es interessante Optionen der Kritik: Kann gezeigt werden, dass Wertfreiheit im Prinzip, also auch für eine ideale Wissenschaftlerin oder einen idealen Wissenschaftler, unmöglich oder nicht erstrebenswert ist, dann wäre das Wertfreiheitsideal auf einer besonders tiefen konzeptionellen Ebene widerlegt. Eine philosophisch anspruchsvolle Kritik sollte hiervon ihren Ausgang nehmen (s. hierzu Teil III).

Danksagung

Dieser Aufsatz wurde im Rahmen des GRK 2073 „Integrating Ethics and Epistemology of Scientific Research" verfasst und von der Deutschen Forschungsgemeinschaft (DFG) finanziert (254954344/ GRK2073). Ich danke Torsten Wilholt und Cornelis Menke für hilfreiche Kommentare zu einer früheren Version. Besonders danke ich Angelo D'Abundo für ein überaus sorgfältiges Korrektorat.

Literatur

Anderson, E. (2004). Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce. *Hypathia*, 19(1), 1–24. doi:10.1111/j.1527-2001. 2004.tb01266.x

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, 3(2), 207–220. https://doi.org/10.1007/s13194-012-0062-x

Biddle, J. (2013). State of the field: Transient underdetermination and values in science. *Studies in History and Philosophy of Science*, 44(1), 124–133. https://doi.org/10.1016/j.shpsa.2012.09.003

Biddle, J., & Winsberg, E. (2010). Value Judgements and the Estimation of Uncertainty in Climate Modelling. In P.D. Magnus, & J. Busch (Hrsg.), *New Waves in Philosophy of Science* (S. 172–197). Palgrave Macmillan.

Bright, L. (2018). Du Bois' democratic defence of the value free ideal. *Synthese*, 95(5), 2227–2245. https://doi.org/10.1007/s11229-017-1333-z

Brown, J. R. (2013). Die Wissenschaftsgemeinschaft – The Community of Science®. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 337–373). Suhrkamp.

Büter, A. (2012). *Das Wertfreiheitsideal in der Sozialen Erkenntnistheorie. Objektivität, Pluralismus und das Beispiel Frauengesundheitsforschung*. Ontos.

Bueter, A. (2015). The irreducibility of value-freedom to theory assessment. *Studies in History and Philosophy of Science*, 49, 18–26. https://doi.org/10.1016/j.shpsa.2014.10.006

Carnap, R. (1950). Empiricism, Semantics, and Ontology. *Revue International de Philosophie*, 4, 20–40.

Collins, H. (1985). *Changing order. Replication and induction in scientific practice*. Sage.

Dahms, H. J. (2013). Bemerkungen zur Geschichte des Werturteilsstreits. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 74–107). Suhrkamp.

De Melo-Martín, I., & Intemann, K. (2016). The Risk of Using Inductive Risk to Challenge the Value-Free Ideal. *Philosophy of Science*, 83, 500–520. https://doi.org/10.1086/687259

Dorato, M. (2004). Epistemic and Nonepistemic Values in Science. In P. Machamer, & G. Wolters (Hrsg.), *Science, Values, and Objectivity* (S. 52–77). University of Pittsburgh Press.

Douglas, H. E. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, 67(4), 559–579. https://doi.org/10.1086/392855

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Dupré, J. (2007). Fact and Value. In H. Kincaid, J. Dupré, & A. Wylie (Hrsg.), *Value-Free Science? Ideals and Illusions* (S. 27–41). Oxford University Press.

Edenhofer, O., & Kowarsch, M. (2015). Cartography of pathways: A new model for environmental policy assessments. *Environmental Science & Policy*, 51, 56–64. https://doi.org/10.1016/j.envsci. 2015.03.017

Falk, S., Rehfeld, D., Römmele, A., & Thunert, M. (Hrsg.). (2006). *Handbuch Politikberatung*. VS Verlag für Sozialwissenschaften.

Feyerabend, P. K. (1975). *Against method. Outline of an anarchistic theory of knowledge*. NLB.

Frisch, M. (2013). Modeling Climate Policies: A Critical Look at Integrated Assessment Models. *Philosophy & Technology*, 26, 117–137. https://doi.org/10.1007/s13347-013-0099-6

Grelling, K., & Nelson, L. (1908). Bemerkungen zu den Paradoxien von Russell und Burali-Forti. In *Abhandlungen der Fries'schen Schule. Neue Folge 2* (S. 301–334). Göttingen.

Haack, S. (2003). Knowledge and Propaganda. Reflections of an Old Feminist. In C. Pinnick, N. Koertge, & R. Almeder (Hrsg.), *Scrutinizing Feminist Epistemology. An Examination of Gender in Science* (S. 7–19). Rutgers University Press.

Hempel, C. (1965). Science and Human Values. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science* (S. 81–96). The Free Press.

Holman, B., & Wilholt, T. (2022). The New Demarcation Problem. *Studies in History and Philosophy of Science*, 91, 211–220. https://doi.org/10.1016/j.shpsa.2021.11.011

Hoyningen-Huene, P. (2006). Context of Discovery Versus Context of Justification and Thomas Kuhn. In J. Schickore, & F. Steinle (Hrsg.), *Revisiting Discovery and Justification. Historical and philosophical perspectives on the context distinction* (S. 119–131). Springer.

Hulme, M. (2009). *Why we disagree about climate change. Understanding controversy, inaction and opportunity*. Cambridge University Press.

Jacobs, R. A. (1985). Is "Ought Implies Can" a Moral Principle? *Southwest Philosophy Review*, 2, 43–54. https://doi.org/10.5840/swphilreview198524

John, S. (2018). Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. *Social Epistemology*, 32(2), 75–87, https://doi.org/10.1080/ 02691728.2017.1410864

John, S. (2019). Science, truth and dictatorship: Wishful thinking or wishful speaking?. *Studies in History and Philosophy of Science*, 78, 64–72. https://doi.org/10.1016/j.shpsa.2018.12.003

Kitcher, P. (2011). *Science in a Democratic Society*. Prometheus Books.

Koertge, N. (2013). Wissenschaft, Werte und die Werte der Wissenschaft. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 233–251). Suhrkamp.

Kourany, J. A. (2003). A Philosophy of Science for the Twenty-First Century. *Philosophy of Science*, 70(1), 1–14. https://doi.org/10.1086/367864

Kourany, J. (2008). Replacing the Ideal of Value-Free Science. In M. Carrier, D. Howard, & J. Kourany (Hrsg.), *The Challenge of the Social and the Pressure of Practice. Science and Values Revisited* (87–109). University of Pittsburgh Press.

Kuhn, T. S. (1977). Objectivity, Value Judgement, and Theory Choice. In *The Essential Tension. Selected Studies in Scientific Tradition and Change* (S. 320–339). University of Chicago Press.

Kuhn, T. S. (1980). *Die kopernikanische Revolution*. Vieweg. (Erstveröffentlichung 1957)

Kühler, M. (2013). *Sollen ohne Können? Über Sinn und Geltung nicht erfüllbarer Sollensansprüche*. Mentis.

Lacey, H. (1999). *Is Science Value Free? Values and Scientific Understanding*. Routledge.

Latour, B., & Woolgar, S. (1979). *Laboratory life. The social construction of scientific facts*. Sage Publications.

Laudan, L. (2004). The epistemic, the cognitive, and the social. In P. Machamer, & G. Wolters (Hrsg.), *Science, values, and objectivity* (S. 14–23). University of Pittsburgh Press.

Leggewie, C. (2006). Deliberative Demokratie – Von der Politik- zur Gesellschaftsberatung (und zurück). In S. Falk, D. Rehfeld, A. Römmele, & M. Thunert (Hrsg.), *Handbuch Politikberatung* (S. 152–160). VS Verlag für Sozialwissenschaften.

Letherby, G., Scott, J., & Williams, M. (2013). *Objectivity and Subjectivity in Social Research*. Sage.

Lipsey, R. G., & Lancaster, K. (1956). The General Theory of Second Best. *The Review of Economic Studies*, 24(1), 11–32

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. E. (1996). Cognitive and Non-cognitive Values in Science: Rethinking the Dichotomy. In L. H. Nelson, & J. Nelson (Hrsg.), *Feminism, Science, and the Philosophy of Science* (S. 39–58). Kluwer.

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Longino, H. E. (2004): How Values Can Be Good for Science. In P. Machamer, & G. Wolters (Hrsg.), *Science, Values, and Objectivity* (S. 127–142). University of Pittsburgh Press.

Longino, H. E. (2008). Values, Heuristics, and the Politics of Knowledge. In M. Carrier, D. Howard, & J. Kourany (Hrsg.), *The Challenge of the Social and the Pressure of the Practice* (S. 68–86). University of Pittsburgh Press.

Luhmann, N. (1971). Die Praxis der Theorie. In *Soziologische Aufklärung. Aufsätze zur Theorie sozialer Systeme* (S. 253–267). Westdeutscher Verlag.

Maasen, S., & Weingart, P. (2005). (Hrsg.). *Democratization of Expertise?. Exploring Novel Forms of Scientific Advice in Political Decision-Making*. Springer.

Machamer, P., & Douglas, H. (1999). Cognitive and social values. *Science & Education*, 8(1), 45–54.

Martin, G. P., Hanna, E., McCartney, M., & Dingwall, R. (2020). Science, society, and policy in the face of uncertainty: reflections on the debate around face coverings for the public during COVID-19. *Critical Public Health*, 30(5), 501–508. https://doi.org/10.1080/09581596.2020.1797997

McMullin, E. (1982). Values In Science. *Proceedings of the Biennial Meeting of the Philosophy of Science Association. Vol. Two: Symposia and Invited Papers*, 3–28. https://doi.org/10.1086/psaprocbienmeetp.1982.2.192409

Mowry, B. (1985). From Galen's Theory to William Harvey's Theory: A Case Study in the Rationality of Scientific Theory Change. *Studies in History and Philosophy of Science*, 16(1), 49–82. https://doi.org/10.1016/0039-3681(85)90007-X

Parker, W. (2014). Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science*, 46, 24–30. http://dx.doi.org/10.1016/j.shpsa.2013.11.003

Pielke, R. A. (2012). *The honest broker. Making sense of science in policy and politics* (8. Aufl.). Cambridge University Press.

Popper, K. (1974). Die Logik der Sozialwissenschaften. In T. Adorno, H. Albert, R. Dahrendorf, J. Habermas, H. Pilot, & K. Popper (Hrsg.), *Der Positivismusstreit in der deutschen Soziologie* (S. 103–123). Luchterhand.

Proctor, R. (1991). *Value-free science? Purity and power in modern knowledge*. Harvard University Press.

Putnam, H. (2002). *The Collapse of the Fact/value Dichotomy and Other Essays*. Harvard University Press.

Reichenbach, H. (1961). *Experience and Prediction*. University of Chicago Press. (Erstveröffentlichung 1938)

Reiss, J., & Sprenger, J. (2020). Scientific Objectivity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). https://plato.stanford.edu/entries/scientific-objectivity/

Rudner, R. (1953). The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science*, 20(1), 1–6. https://doi.org/10.1086/287231

Ruphy, S. (2006). Empiricism all the way down: a defense of the value-neutrality of science in re-sponse to Helen Longino's contextual empiricism. *Perspectives on Science*, 14(2), 189–214. https://doi.org/10.1162/posc.2006.14.2.189

Schickore, J., & Steinle, F. (2006). Introduction: Revisiting the Context Distinction. In J. Schickore & F. Steinle (Hrsg.), *Revisiting Discovery and Justification. Historical and philosophical perspectives on the context distinction* (S. vii–xix). Springer.

Schurz, G. (2013). Wertneutralität und hypothetische Werturteile in den Wissenschaften. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 305–334). Suhrkamp.

Schurz, G., & Carrier, M. (2013). Einleitung und Übersicht. In G. Schurz, & M. Carrier (Hrsg.), *Werte in den Wissenschaften. Neue Ansätze zum Werturteilsstreit* (S. 7–30). Suhrkamp.

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, 77, 14–34. https://doi.org/10.1086/650206

Steele, K. (2012). The Scientist qua Policy Advisor Makes Value Judgments. *Philosophy of Science*, 79 (5), 893–904. https://doi.org/10.1086/667842

Stern, R. (2004). Does 'Ought' Imply 'Can'? And Did Kant Think It Does? *Utilitas*, 16(1), 42–61. https://doi.org/10.1017/S0953820803001055

Talbott, B. (2016). The Best Argument for "Ought Implies Can" is a Better Argument Against "Ought Implies Can". *Ergo*, 3(14), 377–402. https://doi.org/10.3998/ergo.12405314.0003.014

Weber, M. (1988). Die „Objektivität" sozialwissenschaftlicher und sozialpolitischer Erkenntnis. In *Gesammelte Aufsätze zur Wissenschaftslehre* (7. Aufl., hrsg. v. Johannes Winckelmann) (S. 146–214). Mohr. (Erstveröffentlichung 1904)

Weber, M. (1988). Der Sinn der „Wertfreiheit" der soziologischen und ökonomischen Wissenschaften. In *Gesammelte Aufsätze zur Wissenschaftslehre* (7. Aufl., hrsg. v. Johannes Winckelmann) (S. 489–540). Mohr. (Erstveröffentlichung 1917)

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science*, 40(1), 92–101. https://doi.org/10.1016/j.shpsa.2008.12.005

Wilholt, T. (2013). Epistemic Trust in Science. *British Journal for the Philosophy of Science*, 64, 233–253. https://doi.org//10.1093/bjps/axs007

Zecha, G. (1992). Value-Neutrality and Criticism. *Journal for General Philosophy of*

III. INDUCTIVE RISK: DOES IT REALLY REFUTE VALUE-FREEDOM?*

ZUSAMMENFASSUNG Im zweiten Teil habe ich gezeigt, dass das Wertfreiheitsideal auf einer Strategie der Einschränkung durch Abgrenzung beruht. Im dritten Teil diskutiere ich eine der wichtigsten Kritiken dieses Ideal: das Argument des induktiven Risikos. Der besondere Reiz dieses Arguments liegt in der Aussicht, das Wertfreiheitsideal selbst dann zu widerlegen, wenn dessen Einschränkungen anerkannt werden. So soll Wertfreiheit *im Prinzip*, also selbst aus Sicht einer perfekten Wissenschaftlerin oder eines perfekten Wissenschaftlers nicht realisierbar oder nicht erstrebenswert sein. Im Folgenden prüfe ich diese Ambition anhand eines idealisierten Bayesianischen Entscheidungssettings. Ich zeige, dass das Argument tatsächlich erfolgreich ist, allerdings nur gegen den *Sollensanspruch* des Wertfreiheitsideals und nur bei Entscheidungen zwischen Optionen mit gleichem oder ähnlichem Erwartungsnutzen. Weiterhin zeige ich, wie derartige Entscheidungen wertbeladen und dennoch mit zwei der im ersten Teil diskutierten Argumente – dem Präskriptionsargument und einer (leicht modifizierten) Version des Sein-Sollen-Arguments – kompatibel sein können. Am Ende des Teils argumentiere ich, dass die idealisierte Rekonstruktion der Entscheidungssituation relevant für tatsächliches Forschungshandeln ist. Hieraus leite ich ein Konzept von *epistemischer Legitimität* ab, das Regeln für einen akzeptablen Umgang mit wertbeladener Wissenschaft aufzeigt.

* Dieser Artikel wurde zuerst veröffentlicht als Dressel, M. (2022). Inductive Risk: Does It Really Refute Value-Freedom? *Theoria*, 37(2), 181–207. https://doi.org/10.1387/theoria.22795

1. Introduction

For a long time, philosophers maintained that science can and should be value-free. In recent years, however, this value-free ideal (VFI) dramatically lost support (Biddle, 2013; Douglas, 2015; Elliott & Steel, 2017; Holman & Wilholt, 2022). Many specialists in science and values today reject value-freedom, either as a possibility (Wilholt, 2009; Biddle & Winsberg, 2010; Winsberg, 2012) or as an ideal (Douglas, 2009; Steel, 2016a; Elliott, 2011). One of the most salient arguments in this regard is the argument from inductive risk (AIR). From AIR's perspective, science is a sequence of decisions under risk. When the potential consequences of these decisions reach beyond science, AIR claims, scientists cannot or should not remain value-free. What is more, AIR promises to refute value-freedom *in principle*. As inductive risk is part of "the scientific method as such" (Rudner, 1953, p. 2), the idea goes, VFI fails even for the "perfect scientists – the scientist *qua* scientist" (ibid.).

In this paper, I scrutinize exactly this ambition: does AIR refute value-freedom even under the "VFI-friendly" assumption of a "perfect scientist"? I begin by introducing VFI and its conceptual restrictions. I also present two reasons why value-freedom may be appealing in the first place, the argument from prescription (APr) and the argument from wishful thinking (AWT). In the third section, I introduce AIR and develop a Bayesian decision setting in which an ideal agent – the "perfect scientist"– maximizes the expected utility of a given scientific decision. In the fourth section, I use the idealized setting to scrutinize whether the "perfect scientist" is forced, descriptively speaking, to make non-epistemic value-judgements. I conclude that this is not the case, at least not in a way that refutes VFI. However, AIR is more successful in showing that the "perfect scientist" should, normatively speaking, sometimes use non-epistemic values. In the fifth section, I discuss how this is possible while avoiding prescriptiveness and wishful thinking. I argue that, while APr and AWT rightfully warn against problematic ways of using values, these concerns can be countered by taking certain measures. In the conclusion, I discuss the objection that my idealized setting is irrelevant in practice. I argue that this impression is false and that, quite on the contrary, my idealized approach helps to elaborate something that may be called *epistemic legitimacy*: a set of rules that should govern the use of values not only in an idealized setting, but also in actual science.

2. What is value-freedom and why would we want it?

2.1 VALUE-FREEDOM: DEFINITION AND RESTRICTIONS

The version of the value-free ideal (VFI) that I discuss in this paper comprises a normative and a descriptive claim, as well as four restrictions that clarify the scope of these claims. The first claim, VFI_{norm}, describes value-freedom as a regulative ideal:

VFI_{norm} Scientists *should,* as much as possible, avoid significant non-epistemic value-judgments when making genuinely scientific choices.

While some authors have focused on this normative part (e.g. Dorato, 2004; Douglas, 2009, ch. 3; Betz, 2013; Bueter, 2015), others have treated VFI as partially descriptive. For these authors, VFI not only claims that scientists *should* be value-free, but also that they *can* be value-free (e.g. Biddle & Winsberg, 2010; Biddle, 2013; Reiss & Sprenger, 2020). I here follow this twofold interpretation, mostly because the normative and the descriptive part are connected via ought-implies-can (Kitcher, 2011, p. 31; Biddle, 2013, p. 131). I refer to the descriptive part as VFI_{desc}:

VFI_{desc} Scientists *can,* at least in principle, avoid significant non-epistemic value-judgments when making genuinely scientific choices.

As we shall see later, inductive risk challenges the two claims in different ways. To show this, we need to consider VFI's restrictions (see e.g. Weber, 1904/1949; Reichenbach, 1938/1961; Rudner, 1953; Kuhn, 1977; McMullin, 1982; Lacey, 1999; Dorato, 2004; Douglas, 2009; Reiss & Sprenger, 2020):

$VFI\text{-}R_1$ VFI applies only to *non-epistemic* values.
$VFI\text{-}R_2$ VFI applies only to *genuinely scientific* decisions.
$VFI\text{-}R_3$ VFI applies only to *significant* value-judgments.
$VFI\text{-}R_4$ VFI applies only to value-freedom *in principle*.

Critics have argued that these restrictions, particularly VFI-R_1 and VFI-R_2, are analytically implausible or practically infeasible (e.g. Longino, 1996; Machamer & Douglas, 1999; Putnam, 2002; Dupré, 2007; Bueter, 2015; De Melo-Martin & Intemann, 2016). For the scope of this paper, however, I accept these restrictions. I do so not because I think that critiques of VFI-R_1–VFI-R_4 are irrelevant, but because I think that *inductive risk is philosophically interesting exactly because it challenges value-freedom even under assumptions that are favorable for value-freedom*. That is, if AIR defeats VFI even if we stipulate "VFI-friendly" (and perhaps counter-factual) conditions, then this would emphasize the gravity of inductive risk. Hence, while I do not engage in debates about VFI-R_1–VFI-R_4 here, I do contend that these restrictions are useful to study the strengths – and the limits – of inductive risk.

Let us therefore look into these restrictions. VFI-R_1 limits VFI to *non-epistemic* (e.g. ethical) values, but permits *epistemic* values such as explanatory power, scientific fruitfulness, or other values that may foster the attainment of truth (Kuhn, 1977; Steel, 2010). Later in this paper, I will refer to such values as *scientific utilities* (Hempel, 1981). A scientific choice will, e.g., have a high scientific utility if it unifies a research field or enables new lines of study, and a low scientific utility if it leads scientists to accept falsehoods or miss important truths (Steel, 2016a; Wilholt, 2016). The next restriction, VFI-R_2, limits VFI to the "heart of the research process" (Douglas, 2009, p. 45), i.e. to those activities that justify a research finding *qua* truth claim (Weber, 1904/1949; Reichenbach, 1938/1961; Hoyningen-Huene, 2006). This includes activities such as hypothesis assessment, model choice or data collection, but excludes obviously value-laden parts such as agenda-setting, real-world applications and ethical boundary conditions (Douglas, 2009; Biddle, 2013; Reiss & Sprenger, 2020). VFI-R_3 limits VFI to decisions that *significantly* impact final results, e.g. by turning a hypothesis acceptance into a rejection. This makes sure that, if we are to refute VFI, we need an argument that shows how values make an actual difference (rather than an argument that "exaggerates the influence of social values", Parker, 2014, p. 27).

VFI-R_4 is crucial for my analysis of inductive risk. This restriction implies that VFI_{desc} cannot be refuted by showing that scientists *de facto* fail to be value-free, but only by showing that they cannot be value-free *in principle*. I assume VFI-R_4 for three reasons. First, it played a crucial role when inductive risk was originally introduced (Rudner 1953). Second, VFI-R_4 mirrors parts of the recent debate: Reiss & Sprenger (2020) describe VFI_{desc} as the claim that scientists can "at least in principle" (ibid.) refrain from making value judgements; Kitcher (2011) takes proponents of value-freedom to claim

that scientists should "in principle" (ibid., p. 33) report only the evidence; and Ruphy (2006) holds that philosophically interesting critiques of value-freedom are "not only about what doesn't happen *in practice* in science, [but] about what cannot happen even *in principle*" (ibid., p. 192, orig. italics). Third, the *de facto* value-ladenness of science has long been conceded by defenders of VFI. The point is that "[p]roponents of [VFI] may grant that perfectly value-neutral results are never or very rarely obtained in the actual development of science, for all that, value-neutrality remains the aim" (Ruphy 2006, p. 192) (see also Weber 1904/1949, p. 9; Popper 1976, p. 97; Koertge 2000, p. S53). Yet this classic defense presupposes that value-freedom can be achieved at least under idealized assumptions. The question, then, is whether inductive risk really shows that value-freedom fails in principle.

2.2 VALUE-FREEDOM: UNDERLYING MOTIVATION

But why should we be interested in value-freedom in the first place? In this context, Holman & Wilholt (2022) use the metaphor of "Weber's fence". They argue that champions of value-freedom such as Max Weber had relevant reasons to set up a rule, or a "fence", that shields science from inacceptable value influences, and that we should not tear down this "fence" without considering the concerns behind it (see also Proctor, 1991). De Melo-Martin & Intemann (2016) discuss two such motivations: the political concern that "the use of contextual values in scientific reasoning allows scientists to impose their personal value judgments on others" (ibid., p. 503); and the epistemic concern that value-laden science may imply that scientists "accept theories about 'how they wished the world to be' rather than 'how the world really is'" (ibid., p. 502). I refer to the first concern as argument from prescription (APr) and the second as the argument from wishful thinking (AWT).

Versions of APr have been discussed by Weber (1919/1958), Du Bois (1935), and more recently Bright (2018), Betz (2013), De Melo-Martin & Intemann (2016), Intemann (2015), or John (2015, 2019). One way to reconstruct APr goes like this: real-world decision-making often relies on science, be it on an individual level, e.g. regarding a person's medical choices, or on a collective level, say in climate policy or substance regulation. The worry is that, if science would depend on non-epistemic values, scientists would effectively prescribe these decisions. This may violate democratic norms: "to the extent that scientists make value judgments, there are concerns that their values will be undemocratically privileged" (Intemann, 2015, p. 218). The concern can also be expressed as a matter of subjective freedom. Here, the worry is that individuals should be able to

freely pursue their version of the good life, and that, to the degree that they depend on science to do so, "personal autonomy would be jeopardized if scientific findings […] were soaked with moral assumptions" (Betz 2013, p. 207).

Elaborating this a bit further, I suggest that APr is the claim that value-laden science places an illegitimate constraint on the decision space of extra-scientific agents. I suggest that such a constraint is illegitimate if and only if the constraint is *relevant* (eliminating decision options that the agents may actually take interest in), *external* (lacking the agents' explicit or implicit consent) and *normative* (resulting from scientists' non-epistemic values). The underlying principle is best described as the norm of autonomy, which I take to include both democratic and personal autonomy. APr then reads:

1. Extra-scientific agents often rely on science to determine their action plans.
2. Value-laden science constrains the decision space of those agents in a *relevant*, *external* and *normative* way.
3. Such constraints violate the principle of autonomy.
4. Observing VFI_{norm} is the only (or at least the best) available way to avoid this violation.
5. Therefore, VFI_{norm} is valid.

Proceeding to AWT, precursors of this concern can be traced back to Bacon's *Novum Organum* (book I, § 39–46) or Hume's *Treatise of Human Nature* (book III, part I, sect. I). In the recent debate, AWT has been addressed both by critics (e.g. Douglas, 2009, ch. 3; Brown, 2013; De Melo-Martin & Intemann, 2016) and defenders (e.g. Koertge, 2000; Haack, 2003) of value-freedom. AWT claims that "propositions about what states of affairs are *desirable* or *deplorable* [cannot] be evidence that things *are*, or *are not*, so" (Haack, 2003, p. 13, orig. italics). As science is to inform us about actual rather than desirable states of affairs, it seems to follow that science should be value-free. This reasoning stems from a principle called *no-is-from-ought*, the "mirror image" (Jones, 1999, p. 894) of Hume's famous law. Interestingly, however, current discussions of AWT rarely take up insights from meta-ethics (Schurz, 1997; Pidgen, 2016). Closing this gap, I suggest that AWT claims that an *ought-is* inference is logically invalid if and only if its descriptive conclusion is *direct* (derived exclusively from normative premises), *non-vacuous* (a substantial implication of the normative premises) and *not semantically entailed* (not hidden the semantics of the normative premises). AWT thus reads:

1. Scientific choices should be truth-conducive.
2. Making a scientific choice dependent on non-epistemic values is a *direct*, *non-vacuous* and *not semantically entailed* inference from an ought-claim to an is-claim.
3. Inferences of such kind violate no-is-from-ought.
4. Observing VFI_{norm} is the only (or at least the best) available way to avoid this violation.
5. Therefore, VFI_{norm} is valid.

Now, critics of value-freedom can either reject APr and AWT straightforwardly, or they can accept these arguments in general, but claim that they apply only to *some* rather than *all* ways of using values in science. It is my impression that the second strategy reflects the standard view regarding APr, whereas critics are split regarding AWT. Some seem to claim that this concern is somewhat exaggerated (e.g. Brown, 2013; De Melo-Martin & Intemann, 2016), whereas others hold that wishful thinking is a real problem if not properly addressed (e.g. Douglas, 2009; Wilholt, 2009). In this paper, I will only consider the latter strategy. I do so because, for one, I believe that AWT has a strong *prima facie* plausibility and, for another, those who reject AWT tend to rely on arguments other than AIR (De Melo-Martin & Intemann, 2016). However, as my aim is to scrutinize inductive risk rather than to find alternative ways to attack VFI, I will not engage with these debates.

3. Inductive risk challenges value-freedom

3.1 THE ARGUMENT FROM INDUCTIVE RISK

Let us now discuss the *argument from inductive risk* (AIR). Apart from early precursors in scholasticism (Schuessler, 2019) or Blaise Pascal (see also James, 1912), AIR emerged in the middle of the past century (Churchman, 1948; Rudner, 1953; Hempel, 1965) and was later reintroduced by Heather Douglas (2000, 2009). The argument's *locus classicus* is a short article by Richard Rudner (1953). Rudner started by expressing his discontent with popular critiques of value-freedom. These critiques, he claimed, either argue that truth is itself a value, or that values are needed in scientific agenda-setting, or that scientists are imperfect human beings (ibid., p. 1). We can easily see that, unless one rejects VFI's restrictions, none of this refutes value-freedom. In particular, the fact that science

is *de facto* value-laden does not imply that this must be so *in principle* (VFI-R_4). Rudner found this unsatisfying, for as long as values "have not been shown to be involved in the scientific method as such" (ibid., p. 2), it still stands that "[t]he perfect scientist – the scientist *qua* scientist – does not allow this kind of value judgment" (ibid.). Rudner acknowledged that such a "perfect scientist" was nowhere to be found in reality (ibid., p. 4); yet he believed that this idealization was the adequate touchstone for his argument. Hence, Rudner's question (which is also the starting point of my own approach, see sect. 3.2), was essentially counter-factual: *if there were such a thing as an ideal epistemic agent, would this agent make non-epistemic judgements when making genuinely scientific decisions?*

Rudner believed that the answer to this question is "yes". He argued that the scientific method intrinsically involves that "the scientist as scientist accepts or rejects hypotheses" (ibid., p. 2). However, "no scientific hypothesis is ever completely verified" (ibid.). As there is no certainty in science, and as there is no science without hypothesis evaluation, "the scientist must make the decision that the evidence is *sufficiently* strong" (ibid.) before accepting a hypothesis. Rudner argued that the only way to determine "how strong is 'strong enough'" (ibid.) is to weigh the potential consequences of error. When the consequences concern extra-scientific goods, say public health, values are needed to determine how much evidence is needed. Therefore, Rudner argued, hypothesis evaluation is "a function of the *importance*, in the typically ethical sense, of making a mistake" (ibid., orig. italics).

Rudner's reasoning has later been refined and generalized in several ways. First, recent contributions tend to differentiate between a normative version of AIR, i.e. one that attacks VFI_{norm}, and a descriptive version, i.e. one that attacks VFI_{desc} (Betz, 2013; Steel, 2016a)[1]. Second, today's versions typically address not only hypothesis assessment, but also other genuinely scientific activities such as model choice or data interpretation (Douglas, 2000; Elliott, 2011; Wilholt, 2013; Biddle & Kukla, 2017). And third, some current interpretations of AIR address not only the consequences of erroneous scientific decision, but also of correct or suspended decisions (Wilholt, 2009; Wilholt, 2013; Steel, 2016a; Steel, 2016b). Taking up these developments, I interpret AIR in such a way that it includes both a normative and descriptive branch; also, I take AIR to address any genuinely scientific decision that significantly impacts the final results of a scientific

1 De Melo-Martin & Intemann (2016) suggest that AIR may claim that non-epistemic values are necessary in a *logical*, *epistemic*, *pragmatic*, or *ethical* sense. As far as I can see, the first two readings are instances of the descriptive reading, while the latter two are instances of the normative reading.

study (similar to Biddle & Kukla, 2017)[2]; finally, I take AIR to address not only the consequences of error, but also those of truth, missed truth and averted error (similar to Wilholt, 2009) (see sect. 3.2).

AIR, then, is the claim that scientific decisions made by an ideal epistemic agent must, or should, include non-epistemic values if and only if these decisions are underdetermined (involving relevant uncertainty), unavoidable (forced upon the agent) and momentous (having potential consequences for ethically relevant extra-scientific goods). AIR thus reads:

1. An ideal epistemic agent *cannot avoid* making *underdetermined* scientific choices.
2. To make these choices, the agent must specify *evidential thresholds*.
3. If the choice is *momentous* via its potential consequences for ethically relevant extra-scientific goods, the agent *cannot* or *should not* (or both) specify the evidential threshold without considering non-epistemic values.
4. Therefore, if underdetermination, unavoidability and extra-scientific momentousness are given, VFI_{desc} or VFI_{norm} (or both) is (or are) invalid.

3.2 INDUCTIVE RISK IN AN IDEALIZED SETTING

Before discussing whether AIR really defeats value-freedom and, if so, whether AIR can avoid prescriptiveness and wishful thinking, I want to suggest an idealized decision-theoretical approach to inductive risk[3]. My motivation is twofold: On the one hand, I take seriously Rudner's claim that AIR is not merely about actual scientists, but about an idealization – the "perfect scientist" (or "scientist *qua* scientist")[4]. As Katie Steele has

2 Biddle & Kukla (2017) suggest to substitute the term "inductive" risk with "epistemic" risk. I agree that this terminology has virtues. However, I stick to the traditional term because it is commonly used in the debate, and because scientific choices that occur previous to hypotheses assessment (model choice, data collection, test calibration etc.) have basically one purpose: to make possible an inductive step from the evidence to a hypothesis acceptance/rejection (or suspension).

3 The following part is inspired by Wilholt (2009, 2013) and was significantly improved in discussions with Benjamin Blanz and Hermann Held.

4 This is sometimes overlooked. Rudner explicitly says that his considerations "do not have as their import that an empirical description of every present day scientist […] would include the statement that he made a value judgment" (1953, p. 4). Rudner's point was rather that a "rational reconstruction of the method of science" (ibid.) would be incomplete if it did not address inductive risk. While Rudner noted that scientists are not "coldblooded,

pointed out, such counter-factual assumptions would strengthen AIR, "because it is more surprising in the ideal setting that scientists must make value judgments" (2012, p. 895). Also, an idealized approach provides an *in-principle* perspective on value-freedom (as demanded by VFI-R_4). On the other hand, the approach sheds light on decision problems in *actual* science. Inductive risk has often been described as a balancing problem between exactly two risks, where one risk is clearly preferable to the other (e.g. consumer versus producer risks, Biddle & Leuschner, p. 2015). As shown below, however, the decision problem is much more complex.

To represent Rudner's "perfect scientist", I stipulate an ideal agent with the following properties:

* *Preferences.* The agent prioritizes the advancement of science over extra-scientific aims.
* *Evidence.* The agent possesses perfect knowledge of the available evidence.
* *Rationality.* The agent makes decisions in a rule-based and unbiased manner.

Before this background, the agent considers a scientific decision *D*, where *D* may be any methodological choice that significantly impacts the final results of the study that *D* is a part of. As an example, imagine that the agent contemplates whether or not to use a certain model. The agent's decision space comprises two options: *perform D* (use the model) and *not perform D* (not use the model)[5]. A central assumption of AIR is that agent cannot be certain whether performing *D* (using the model) would lead to true study results[6]. The agent must therefore determine a threshold *t* above which the

emotionless, impersonal" (ibid, p. 6), he came to this conclusion not by considering actual science, but by analyzing an *impersonal* scientist qua scientist.

5 For reasons of simplicity, I focus on: (a) individual decisions rather than decision sequences; (b) binary decisions (e.g. use versus not use a model) rather than decisions with three options (e.g. accept, reject or suspend a hypothesis); and (c) decisions on single methodological items (e.g. a model) rather than contrastive decisions between different items (e.g. several competing models). Note, however, that my approach could in principle accommodate these types of decisions.

6 The extent to which scientists can avoid uncertainty is contested (Betz, 2013; Parker, 2014; Steel, 2016a; Douglas, 2017). Note, however, that AIR need not assume that each and every scientific choice is fundamentally uncertain. In fact, I believe that this radical interpretation of AIR is either trivial or false. There must be a difference between the trivial uncertainty attached to, say, the assumption that radiative forcing is a relevant factor in the climate system, and the non-trivial uncertainty attached to, say, cloud parametrizations in a given climate model. However, for inductive risk to be relevant it suffices that non-trivial uncertainty is a typical feature of science,

probability p that the results will be true, given that D is performed (the model is used), is sufficiently high. The agent would then perform D when p exceeds t, and not perform D when p falls short of t:

Perform D iff $p > t$
Not perform D iff $p < t$

The question, then, is how the agent determines the evidential threshold t. The classic answer (Rudner, 1953; Churchman, 1948) is that the threshold depends on how bad the consequences of error would be. However, this leaves open many critical issues: How exactly do D's consequences determine t? How should scientific and extra-scientific consequences be balanced? How should outcomes other than error influence t? What is the relation between the probability that D leads to an error and the probability that the error causes the assumed consequences? Another issue is that the classic approach interprets inductive risk in a frequentist manner. In frequentist statistics, p is an objective measure for the likelihood with which a property that has been found in a number of observations $O_1,\ldots,O_n$ will also be found in an observation O_{n+1} (Rudner, 1953, p. 3). Our agent, however, is in a different epistemic situation: the available evidence may be too limited or inconsistent to determine an objective probability; the evidence may be incommensurate, e.g. because it includes data from heterogeneous sources; and the evidence does not, by definition, account for unknown unknowns. It is therefore more plausible to interpret p in a Bayesian manner, such that p represents the agent's *probabilistic beliefs* (see also Steele, 2012).

By adopting a Bayesian perspective, I also contend that the agent can be understood as a *utility maximizer*[7]. That is, the agent will choose the option that promises the highest

and that in a relevant number of cases this uncertainty cannot be avoided without sacrificing science's ability to produce meaningful results. This more modest reading of AIR accounts for the possibility of uncertainty hedging (Betz, 2013), while still reserving a crucial role for inductive risk.

7 Note that I do not claim that utility maximization is the only plausible candidate for a rational decision rule. What I do contend, however, is that the Bayesian perspective is superior to both the frequentist approach and the simplistic decision rule "the worse the error consequences, the higher the evidential threshold" (for reasons outlined above). Note furthermore that the Bayesian approach does not contradict the fact that scientists typically put special emphasis on error avoidance (Wilholt, 2009), as this can easily represented by asymmetrically decreasing the utility of error consequences. Finally, the Bayesian approach can be reconciled with the deontologist axiom that some decisions are intrinsically inacceptable. As I argue later (sect. 5.2), however, such cases should be interpreted as ethical rather than genuinely scientific choices.

Table 1: The agent's decision algorithm.

Step	Description	Formalization
(1)	Determine the probability P of all second-order outcomes (2ndOrd), given the respective first-order outcome (1stOrd). First-order outcomes include truth, error, averted error, and missed truth; second-order outcomes include scientific (2ndOrdSci) and extra-scientific (2ndOrdExt) outcomes. Each first-order outcome may cause several scientific and several extra-scientific second-order outcomes. Extra-scientific second-order outcomes are generally uncertain; scientific second-order outcomes are either certain or uncertain (see fn. 8).	$1 \geq P(2ndOrdSci_i \mid 1stOrd_k) \geq 0$ $1 > P(2ndOrdExt_j \mid 1stOrd_k) > 0$ $i \in \{1, \ldots, n\}$ $j \in \{1, \ldots, m\}$ $k \in$ {Truth, Error, AvertedError, MissedTruth}
(2)	Determine the expected utility EU_{ind} of all individual second-order outcomes, given the respective first-order outcome, based on the second-order outcomes' dependent probability P and the second-order outcomes' utility U.	$EU_{ind}(2ndOrdSci_i \mid 1stOrd_k) = U(2ndOrdSci_i) \cdot P(2ndOrdSci_i \mid 1stOrd_k) + U(\neg 2ndOrdSci_i) \cdot P(\neg 2ndOrdSci_i \mid 1stOrd_k)$ $EU_{ind}(2ndOrdExt_j \mid 1stOrd_k) = U(2ndOrdExt_j) \cdot P(2ndOrdExt_j \mid 1stOrd_k) + U(\neg 2ndOrdExt_j) \cdot P(\neg 2ndOrdExtj \mid 1stOrd_k)$
(3)	Aggregate the individual expected utilities into the aggregated expected utility EU_{agg} of the scientific and extra-scientific second-order outcomes for each first-order outcome.	$EU_{agg}(2ndOrdSci \mid 1stOrd_k) = \sum_{i=1}^{n} EU_{ind}(2ndOrdSci_i \mid 1stOrd_k)$ $EU_{agg}(2ndOrdExt \mid 1stOrd_k) = \sum_{j=1}^{m} EU_{ind}(2ndOrdExt_j \mid 1stOrd_k)$
(4)	Determine a setting for the trade-off parameter T.	Use only scientific utilities iff $T = 1$ Use only extrascientific utilities iff $T = 0$ Use both scientific and extrascientific utilities iff $1 > T > 0$
(5)	Determine the T-weighted expected utility EU_{T} of each first-order outcome.	$EU_T(1stOrd_k) = EU_{agg}(2ndOrdSci \mid 1stOrd_k) \cdot T + EU_{agg}(2ndOrdExt \mid 1stOrd_k) \cdot (1 - T)$
(6)	Determine the probability p that performing D (PerfD) will imply true results.	$1 > p > 0$ $p = P(Truth \mid PerfD) = P(MissedTruth \mid \neg PerfD)$ $1 - p = P(Error \mid PerfD) = P(AvertedError \mid \neg PerfD)$
(7)	Determine the total expected utility EU_{total} of the decision options perform D (PerfD) and not perform D (¬PerfD), given p.	$EU_{total}(PerfD) = EU_T(Truth) \cdot p + EU_T(Error) \cdot (1 - p)$ $EU_{total}(\neg PerfD) = EU_T(MissedTruth) \cdot p + EU_T(AvertedError) \cdot (1 - p)$
(8)	Determine the evidential threshold t.	$t := EU_{total}(PerfD) = EU_{total}(\neg PerfD)$ $t = \frac{EU_T(AvertedError) - EU_T(Error)}{EU_T(Averted\ Error) - EU_T(Error) + EU_T(Truth) - EU_T(MissedTruth)}$

relative benefit, given her preferences regarding the consequences and the probability that she assumes for these consequences to occur (Wilholt, 2009; Wilholt, 2013). Apart from addressing the above issues, this sheds light on the old problem (Kuhn, 1977) that epistemic values such as precision and scope can contradict each other. From a Bayesian perspective, it is irrelevant whether, say, a model's strengths in precision are countered by its weaknesses in scope, as this simply reduces the model's overall utility. The most crucial advantage, however, is that the Bayesian approach gives us a straightforward interpretation of the evidential threshold *t*, where *t is the point in the probability space at which the total expected utility* EU_{total} *of both decision options, i.e. perform* D *(PerfD) and not perform* D *(¬PerfD), converge*:

$$t := EU_{total}(Perf\ D) = EU_{total}(\neg PerfD)$$

To determine *t*, the agent must thus determine both options' total expected utilities. I here suggest to differentiate between *first-order* and *second-order* outcomes (see fig. 1). Drawing on Wilholt (2009, 2013), first-order outcomes include *truth* and *error* for performing *D*, and *missed truth* and *averted error* for not performing *D*. For instance, the agent may correctly decide to use a model that leads to valid study results (truth); erroneously decide to use a model that leads to false study results (error); erroneously decide not to use a model that would have led to valid study results (missed truth); or correctly decide not to use a model that would have led to false study results (averted error). Second-order outcomes are dependent on first-order outcomes, i.e. they may occur as a causal effect of truth, error, missed truth, or averted error. Second-order outcomes include all normatively relevant consequences that *D* may have for both scientific and extra-scientific goods. If the agent, e.g., uses a model that turns out to imply true study results (first-order outcome), this may enable new lines of study (scientific second-order outcome), while also supporting real-world decision-making in, say, climate policy (extra-scientific second-order outcome). The agent must therefore assess how good or bad each second-order outcome would be *if* it occurred, and how likely it is *that* it occurs, given the respective first-order outcome[8].

8 Probabilities in the second-order outcome space are subjective (i.e. they represent the agent's probabilistic beliefs) and dependent (i.e. they are estimated given the respective first-order outcome). Extra-scientific outcomes are generally uncertain, as the agent cannot know whether the study will actually influence real-world contexts. Scientific outcomes are uncertain if they refer to future research (e.g. a result's fruitfulness); however, the agent can be certain about some types of scientific outcomes, such as a result's scope or precision, as compared to existing results.

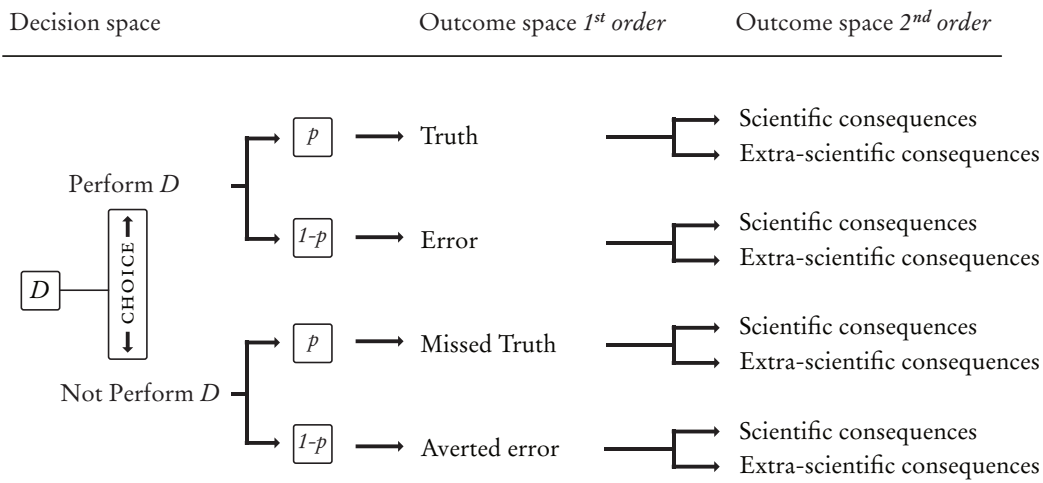


Figure 1: The structure of scientific decisions under inductive risk.

The next step is crucial: the agent must determine the right balance between scientific and extra-scientific utilities. To represent this choice, I suggest to introduce a trade-off parameter T. The agent uses T such that she weighs scientific utilities with T and extra-scientific utilities with $1 - T$. If VFI is valid, the agent will thus use $T = 1$; however, if AIR's attack on value-freedom is successful, the agent must or should (or both) use $T < 1$. The introduction of T gives us a more detailed perspective on AIR. Inductive risk debates have long focused either on *whether or not* non-epistemic values do or should influence scientific decisions, or on the *specific way* in which they should do so. While these questions are indeed crucial, they do not account for the *relative weight* that non-epistemic values should have, as compared to the epistemic ones. Scientific and extra-scientific utilities are different criteria, and merely knowing how high or low a decision option scores in one criterion does not tell us how important the criterion itself is (e.g. how important a scientifically valuable result is in comparison to improved real-world decisions). T represents this balancing problem in a more fine-grained way than, e.g., the notion of one type of value "trumping" (Elliott & McKaughan, 2014) the other.

4. Challenging the challenger: Does inductive risk really refute value-freedom?

4.1 DOES AIR REFUTE VFI_{DESC}?

In this section, I use the technical basis developed above to discuss AIR's challenge of value-freedom. I first discuss whether the agent *can* come to a conclusion about *D* without considering extra-scientific utilities, and then discuss whether, and under what circumstances, the agent *should* use extra-scientific utilities. Starting with the first question, there seems to be an obvious problem with the inductive risk story: if the extent to which extra-scientific utilities influence *D* is regulated by *T*, and if $T = 1$ is a possible setting, then the claim that value-ladenness is inevitable seems to be trivially false. However, there may still be ways in which *D* could be value-laden even under $T = 1$:

(i) Non-epistemic values could be hidden in the scientific utilities.
(ii) Non-epistemic values could be hidden in *p*.
(iii) Non-epistemic values could be hidden in *T*.

As it turns out, all three claims may be true, but not in a sense that threatens VFI_{desc} in our decision setting. Regarding (i) and (ii), a strategic and a substantial point should be considered. The strategic point is that proponents of AIR are not well advised to focus on (i) or (ii), as this would undermine much of AIR's appeal. If it is true that AIR is philosophically interesting because it challenges VFI even under "VFI-friendly" assumptions, supporters of AIR should not claim that values are basically *everywhere*, but rather focus on the specific way in which values influence *D* via the evidential threshold *t*. The trouble is that inductive risk is irrelevant in (i) and (ii), as even a decision that is highly certain and non-risky for extra-scientific goods – e.g. accepting the statement that cold and salty water sinks to deeper ocean layers – will be value-laden if non-epistemic values were hidden in *p* or in scientific utilities such as explanatory power. This does not go well with the inductive risk narrative, which fundamentally depends on *relevant* uncertainty (Betz, 2013) and on *identifiable* causal effects on extra-scientific goods. Also, (i) is effectively a rejection of VFI-R_1 (the restriction that VFI only refers to non-epistemic values). Of course, this does not mean that (i) or (ii) are false; rather, it means that the aspiration to refute VFI even under conditions that are favorable for value-freedom cannot be maintained. From a strategic point of view, those who believe that AIR con-

stitutes one of the strongest challenges of value-freedom should therefore not attempt to capitalize on (i) and (ii).

But there is a more substantial case against (i) and (ii). With respect to (i), we should clarify what exactly the agent maximizes under $T = 1$: *D*'s *conditional* utility for a given understanding of science, or *D*'s *unconditional* utility for some transcendent idea of "science as such"? Clearly, it is the former. Understandings of science change over time, and even at a given point in time there may be more than one definition of good science (just think of the current debate among data scientists and statisticians about whether models should rather be interpretable or accurate, Hassani et al., 2021). As "science as such" is an ill-defined term, the agent can only maximize *D*'s utility for a *given* version of science. *D* will thus have a different utility in a version of science that favors, say, simplicity, than in one that favors heterogeneity (Longino, 2008). But this can be said about *any* decision. For instance, a person who aims to maximize private wealth will make different choices if "wealth" refers only to financial resources than when it also includes resources such as time. And surely, determining whether time counts as wealth is a value-judgement. However, once the general goal is sufficiently defined, judgements about a choice's utility are rather instrumental than genuinely normative. We may thus grant that the *general* aims of science involve non-epistemic judgements (Kitcher, 2001; Bueter, 2015) and still maintain that the agent need not make such judgements in a *specific* decision.

The same is true for (ii). We may grant that *p* depends on the kinds of truths that science is supposed to find; but this does not mean that, once these expectations are set, the agent needs to consider non-epistemic values in any specific decision. For instance, if *p* were to represent the probability with which a climate model will produce trustworthy estimations of climate impacts, then *p* depends *inter alia* on what counts as an "impact" (De Melo-Martín & Intemann, 2016). This, in turn, depends on "judgments about what goods are worth protecting" (ibid., p. 514). However, once this has been determined, the agent need not make any further value-judgement when determining *p*. Hence, even if (i) and (ii) were true, they would not refute $\mathrm{VFI}_{\mathrm{desc}}$ in the decision setting under consideration.

So what about (iii)? We may here think of an argument that shows that $T = 1$ is itself a value-judgement. One promising basis for such an argument could be Jürgen Habermas' (1979, 1998) speech act theory. Every communicative act, Habermas argued, presupposes certain validity claims, one of which he called *moral rightness*. For instance, if we order a non-vegetarian meal in a restaurant, we implicitly claim that eating animals is per-

missible, irrespective of whether we considered this ethical claim in our actual decision process. Yet if a speech act such as "I would like to order the steak" entails "it is morally acceptable to eat animals", then the act of performing or not performing D under $T = 1$ seems to entail "it is morally acceptable to disregard D's effects on extra-scientific goods". As this seems to be a non-epistemic value-judgement, D seems to be value-laden even under $T = 1$.

Now, whether or not we accept this argument depends on what we think the propositional content of the entailed normative claim is. One might argue that the entailed claim is that it is acceptable for scientists to cause D's specific extra-scientific consequences. For instance, if a model choice could affect the regulation of a given toxicant, then $T = 1$ seems to entail that it is permissible to cause exactly the effects that this regulation may have – say, an increase in cancer rates. But this interpretation is misleading. First, $T = 1$ clearly does not presuppose that it is permissible to *cause* such consequences, but rather that it is permissible to *ignore* them. There is a difference between ignoring something and bringing something about. Funding agencies, for instance, may rightfully ignore whether a rejection hampers an applicant's career, but they may not intentionally cause such harm. Second, the entailed claim is *unspecific*, i.e. it refers to *any* extra-scientific consequence. The consequences could therefore change without changing D. This is very different to the restaurant example, where varied consequences can actually change the decision (imagine the choice harmed not cattle but, e.g., dogs). Third, and most importantly, the entailed claim is *redundant*: that it is acceptable to ignore D's extra-scientific consequences simply means that it is acceptable to make value-free scientific choices. Contrary to the restaurant example, this is not an independent proposition, but rather a trivial implication of $\mathrm{VFI}_{\mathrm{norm}}$ – any norm presupposes that it is permissible to observe the norm ("ought" trivially implies "may"). As the claim entailed by $T = 1$ does not contain anything that was not already obvious, this type of value-judgement is uninteresting in our context. We must thus conclude that our agent may indeed presuppose non-epistemic judgements under $T = 1$, but not in a sense that undermines $\mathrm{VFI}_{\mathrm{desc}}$.

4.2 DOES AIR REFUTE $\mathrm{VFI}_{\mathrm{NORM}}$?

The above conclusion is consistent with other authors (Betz, 2013; De Melo-Martin & Intemann, 2016; Steel, 2016a) who, with different arguments, have also claimed that AIR does not refute $\mathrm{VFI}_{\mathrm{desc}}$. Let us therefore see whether AIR is more successful in showing that the ideal agent *should* use non-epistemic values. Regarding this question, we need to consider two cases: one case where $T = 1$ implies equal (or at least very similar) expected

scientific utilities for both decision options (case A); and another case where $T = 1$ implies a relevant difference between the two options, such that either performing or not performing D scores higher in expected scientific utilities (case B):

Case A $EU_{total}(Perf\ D, T=1) = EU_{total}(\neg PerfD, T=1)$
Case B $EU_{total}(Perf\ D, T=1) \neq EU_{total}(\neg PerfD, T=1)$

It turns out that case A provides much stronger grounds for attacking VFI_{norm} than case B. Case A describes a state of *epistemic indifference*, i.e. a situation in which both decision options are equally promising regarding their desired scientific effects. The agent can therefore pursue her primary goal – the advancement of science – equally well by performing or by not performing D. In order to resolve the indifference, the agent has two options at her disposal: either she leaves $T = 1$ unchanged and "simply rolls a die" (Betz, 2013, p. 210); or she decreases the T-setting ($T < 1$) to a level where one decision option scores higher in total expected utilities than the other. In such a situation, it seems obvious that the agent should not randomize the choice, but rather decrease T. The striking reason is that *the surplus in expected extra-scientific utilities does not come at the expense of the expected scientific utilities*. If the agent can maximize both types of utilities at the same time, it is highly implausible that she should jeopardize the extra benefit by rolling a dice. Not only is it *irrational* to reject the raise in total expected utilities, it is also *blameworthy*, as failing to do good when it comes without costs is inappropriate even for an ideal epistemic agent. After all, the agent's commitment to scientific aims does not justify moral indifference, as long as the moral aims are compatible with the primary aim. Situations of epistemic indifference hence constitute a strong case against VFI_{norm}.

The idea that non-epistemic values should work as "tie breakers" to resolve epistemic indifference has been proposed by others (Steel, 2010; Steel & Whyte, 2012; Winsberg, 2012). Yet, it is important to see what exactly this means. "Tie breaker" situations have sometimes been described as "cases where hypotheses score equally well with respect to the evidence" (Magnus, 2018, p. 415, see also Intemann, 2005, p. 1007; Brown, 2013, p. 832). From a decision-theoretical perspective, however, this is only half true. Evidential support, i.e. p, is only *one* parameter that influences the expected scientific utility of a decision option; besides p, the agent must also consider U (the utility of D's consequences) and P (the dependent probability that these consequences actually occur). For instance, if two options are equally well supported by the evidence, but one option scores higher in U and P (e.g. because it will very likely have very positive impacts on future research),

then the expected scientific utilities of the two options will diverge. The agent can therefore have a strong preference despite an identical p (Wilholt, 2009). Hence, contrary to some interpretations of the "tie breaker" thesis, equal evidential support alone does not constitute epistemic indifference. Irrespective of the interpretation, however, the "tie breaker" thesis expresses a valid idea: that even the "perfect scientist" should consider non-epistemic values if she can do so without compromising her scientific preferences.

Some authors, however, argue that non-epistemic values should also be considered in case B, i.e. in a scenario where the agent has a clear epistemic preference (Brown, 2013; Elliott & McKaughan, 2014; Intemann, 2015; De Melo-Martin & Intemann, 2016). While I agree that this may (at least sometimes) be plausible in actual science, I disagree that such an argument can be made for Rudner's "perfect scientist". The problem is that, contrary to case A, adopting $T < 1$ in case B *can* be scientifically detrimental. This can occur when the expected scientific and extra-scientific utilities pull into opposing directions. Imagine a situation where the introduction of a new model may be highly beneficial for the future development of a given research area, e.g. because it eliminates existing inconsistencies or enables new types of questions; yet this model may also make the research field less applicable to real-world problems, e.g. because the model's practical implications are ambiguous or because it generates data that are irrelevant for real-world decisions. It is hard to see why, in such a situation, the "perfect scientist" should disregard the scientific benefits and favor the extra-scientific benefits instead. After all, a crucial part of what it means to be a "perfect scientist" is exactly this: to prioritize the advancement of science. Choosing an option that may be scientifically detrimental is clearly incompatible with this preference. Hence, while AIR is strong in case A, it fails in case B.

Let me now discuss three questions that immediately emerge from the above considerations:

1. I have argued that AIR succeeds only in case A, i.e. in a scenario where the expected scientific utilities of performing and not performing D are identical. However, this scenario seems to be rather untypical. We thus have to ask how relevant AIR's success against $\mathrm{VFI}_{\mathrm{norm}}$ really is.
2. I have argued that AIR does not succeed in case B, as the "perfect scientist" cannot favor extras-scientific over scientific benefits. Yet, this seems to presuppose that scientific and extra-scientific utilities imply opposing decisions. This raises the question how the agent should act when both types of utilities pull into the same (rather than the opposite) direction.

3. I have argued that the agent cannot jeopardize her scientific preferences without ceasing to be a "perfect scientist". At the same time, I have said that this may not necessarily be so in actual science. The question is thus how relevant the above reasoning is for actual science.

I discuss the first two questions here and consider the third question in the conclusion. Regarding the first question, I concede that an exact convergence of expected scientific utilities (case A) may seem untypical, thus creating an impression of irrelevance. Yet, this impression is false. First, even if *exact* convergences were untypical, utilities may well be *approximately* equal. Which option the agent choses would then be rather unimportant for science. Given this lack of significance, we can plausibly treat approximate and exact epistemic indifference analogously, which broadens the set of scenarios covered by case A. Second, there are contexts where epistemic indifference is not uncommon at all, namely when a research field is still young. In avant-garde science, it is often unclear which option will yield higher scientific benefits, as the field's future development is highly uncertain. Third, the impression that epistemic indifference is untypical rests on the assumption that p represents a point prediction. However, as Wendy Parker has argued, "one must know a lot to be a position to say with justification that the probability (degree of belief) that should be assigned to a hypothesis is 0.38 rather than 0.37 or 0.39" (2014, p. 27). Whenever the evidence is scarce, inconsistent, or ambiguous, p will plausibly be expressed as an interval, say [0.3, 0.4] rather than 0.38. Note that this holds even for the "perfect scientist", who is just as confined to the currently available evidence as actual scientists are. Yet, as soon as p comes as an interval, epistemic indifference is more likely. Case A, and hence AIR's refutation of $\mathrm{VFI}_{\mathrm{norm}}$, is thus more relevant than it may seem at first sight.

Regarding the second question, note that case B comprises two different scenarios: one where D is expected to be beneficial for science, but detrimental for extra-scientific goods; and one where D promises scientific and extra-scientific benefits at the same time. Critics of value-freedom tend to focus on the first scenario, where there is a trade-off between scientific and extra-scientific considerations (Douglas, 2000; Douglas, 2009; Elliott & McKaughan, 2014). As noted by Steel (2016b), however, epistemic and non-epistemic values need not necessarily pull into opposite directions. For instance, scientific simplicity can be good for both extra-scientific decision-making (by providing quick results) and for science (by reducing complexity). Interestingly, this non-trade-off scenario is irrelevant and relevant at the same time. It is irrelevant as extra-scientific utilities do not change D if they merely reconfirm an already existing preference. Also, remem-

ber that the version of VFI under consideration is restricted to judgements that actually change a decision, e.g. from using to not using a model (VFI-R_3). Unless one rejects VFI-R_3, it thus follows that the agent is permitted, although not obliged, to consider non-epistemic values in a non-trade-off scenario. Of course, whether or not she does so is effectively irrelevant, at least from a consequentialist perspective (AIR is obviously an instance of consequentialist ethics). Yet, the non-trade-off scenario is relevant in a different sense. Inductive risk narratives can create the impression that there is an intrinsic conflict between doing what is good for science and doing what is good from an ethical perspective. While such conflicts exist, they are clearly contextual, i.e. they may occur or not. The relevance of the non-trade-off scenario is thus that it shows that $T = 1$ need not necessarily imply ethically undesirable results.

5. Can AIR avoid prescription and wishful thinking?

5.1 APR'S CHARGE OF PRESCRIPTIVENESS

The previous section has argued that even the "perfect scientist" should sometimes use non-epistemic values. However, I have also argued that Holman & Wilholt (2022) and others are right to claim that we should not tear down "Weber's fence" (ibid.) without addressing VFI's concerns. In this section, I will thus discuss how *D* can be value-laden, yet not prescriptive and logically fallacious. I start by discussing the argument from prescription (APr)[9]. APr's main claim reads (see sect. 2.2):

APr (2) If *D* is value-laden, it constitutes a *relevant*, *external* and *normative* constraint of extra-scientific agents' decision space (which violates their autonomy).

9 This part benefited from discussions with members of the Consortium for Science, Policy and Outcomes at Arizona State University and the Mercator Research Institute on Global Commons and Climate Change, particularly Martin Kowarsch.

As said before, critics of VFI can either reject this claim, e.g. by arguing that *D* does not really constrain extra-scientific agents[10] or by arguing that such constraints are actually legitimate[11]. Alternatively, they can accept APr in general, but argue that – if the right measures are taken – *D* does not fulfill at least one of APr's conditions (relevance, externality, normativity). As previously said, I only discuss the latter strategy. Two conditions are promising for this strategy: relevance and externality. A constraint is *relevant* if it removes options from an agent's decision space that the agent may actually take interest in; a constraint is *external* if the agent did not, explicitly or implicitly, consent to the constraint. The third condition, normativity, refers to *D*'s value-ladenness. This condition makes sure that those scientific choices that are not value-laden in AIR's sense (e.g. accepting the statement "coal burns", Betz, 2013, p. 21) cannot qualify as prescriptive. However, as we are here interested in cases were *D* includes extra-scientific utilities, the normativity condition is obviously fulfilled.

So what about relevance? To illustrate this condition, consider Rudner's example of the Manhattan Project. Before conducting their detonation experiments, the involved scientists had to accept "the hypothesis that no uncontrollable pervasive chain reaction would occur" (1953, p. 2-3). Assuming that they considered extra-scientific utilities, we can take it that *U* was high for preventing the nuclear accident, and low for causing it. Was this judgement prescriptive? Obviously not. As none of the potentially affected stakeholders can have preferred the accident, the judgement did not restrict anyone's

10 Critics of the so-called "linear model of expertise" (Jasanoff & Wynne, 1998) have argued that "the influence of science on policy is [not] strong and deterministic" (Beck, 2011, p. 298). In their view, actual science-policy processes show that "[i]t would be an exaggeration to state that science [is] driving this process" (Grundmann & Rödder, 2019, p. 4). This may undermine APr's claim that science constrains real-world decisions. But this reasoning is implausible. While both APr and AIR assume that *D* influences extra-scientific agents, neither of them presupposes determinism. The false impression stems from confusing *D*'s first- and second-order outcomes. It then seems that, if *D* implies a true or a false result, certain extra-scientific effects must occur. However, since extra-scientific effects are mediated by various factors (individual reflection, public debate, political compromise etc.), this is clearly false. The notion of "decision constraint" should hence be interpreted probabilistically (via *P*), such that *D* makes it more or less *likely* that extra-scientific agents make certain decisions.

11 APr presupposes some commitment to liberal democracy. However, some authors argue that liberal freedom is less important than other goods such as the prevention of environmental disasters. James Lovelock (2010), e.g., has famously argued that climate change may make it necessary "to put democracy on hold for a while" (ibid.) (see also Shearman & Smith, 2007; Beeson, 2010). Supporters of this reasoning could hence argue that the principle of autonomy is too weak to sustain APr. However, such claims typically presuppose some argument from emergency. Even if such considerations were successful, they would thus undermine APr only in exceptional cases.

freedom of choice[12]. One way to avoid prescriptiveness is thus to use only uncontroversial values. Yet, this approach has limits. More often than not, there will be no consensus on extra-scientific utilities. Even in Rudner's example, the consensus comprises only the consequences of error and averted error, while the extra-scientific effects of truth (building the atomic bomb) and missed truth (not building the atomic bomb) are clearly controversial. Furthermore, scientists may assume a consensus where there is none. This problem can be mitigated, e.g. by conducting stakeholder surveys and by using scenario approaches (Edenhofer & Kowarsch, 2015) that include "solution pathways for any of the major attitudes that can be found in society" (Held 2011, p. 115). Note, however, that this will not always be possible. While, e.g., climate researchers need not commit themselves to only one climate projection, the scientists in Rudner's example could either conduct or not conduct the experiment, but not both. Also, surveys and scenario approaches are again subject to inductive risk (choice of sample sizes, definition of scenarios etc.), thus repeating the prescriptiveness issue on a higher level. Attempting to use only uncontroversial values may thus not always be successful.

Let us therefore consider APr's externality condition. Critics of VFI have offered two strategies to avoid externality, the *transparency* and the *democratic* approach. In the former, scientists determine extra-scientific utilities by themselves, but communicate their choices transparently (Rudner, 1953, p. 6; Douglas, 2009, ch. 4; Elliott & McKaughan, 2014). Extra-scientific agents can then scrutinize these choices and, should they disagree, simply ignore the respective study. This protects their freedom of choice. The problem with this approach is that it views autonomy as an *ex post* capacity, i.e. as the right to reject or accept a choice that has already been made. Call this autonomy *qua* recipient. Moreover, it seems implausible that extra-scientific agents can easily "backtrack" value-judgments, as Elliott & McKaughan (2014) have argued, "and adopt their own alternative assessments and conclusions" (ibid., p. 16). For this to be possible, the implications of these judgements must be deducible just by extrapolation. In most cases, however, extra-scientific agents will only have rough clues what a study would have looked like if, say, a different model would have been used. Thus, while the transparency approach has the virtue of practicality, it promotes only a weak form of autonomy.

12 Stephen John (2019) has recently suggested a notion of "value-aptness" that seems to point into a similar direction (although John refers to the *communication* of scientific findings, not the making of the scientific decision as such). John argues that value-laden communication by scientists does not violate the audience's autonomy if the underlying values are compatible with the values held by the audience. An implication of John's "value-apt ideal" would thus be that the employed values no longer constitute a relevant decision constraint. As discussed below, however, avoiding the relevance condition is only one way to avoid illegitimate prescription.

In contrast, the democratic approach promotes an *ex ante* notion of autonomy, where stakeholders are consulted before the respective judgements are made. Call this autonomy *qua* author. Clearly, being the author of a value-judgement allows for more autonomy than being its recipient. Such authorship may be realized in various ways. The most ambitious forms are iterative (steady consultations rather than one-time interactions), direct (involving ordinary citizens rather than professional representatives), deliberative (consensus-oriented and rational) and inclusive (involving all affected parties) (Douglas, 2005; Douglas, 2009, ch. 8; Brown, 2009, ch. 9-10; Kitcher, 2011; Kowarsch, et. al 2016). Citizen panels are a good approximation to this ideal (Davies et al., 2005; Tomblin et al., 2017). The trouble is that such formats are slow and costly, thus diminishing resources that could be used for other scientific and social projects. Moreover, they may be suited to discuss the general normative issues of a research field, but not the numerous, highly technical decisions that must be made in an individual study. These problems can be mitigated, e.g. by using less iterative or less direct forms of participation. Participation could also be restricted to a higher institutional level, such that extra-scientific agents contribute to the production of general guidelines, but not to their application in specific studies (Steel, 2016a). But this does not come without downsides either. The less stakeholders participate in making the actual judgement, the less can they be seen as its authors; also, the "downscaling" of general guidelines to specific scientific choices will again be subject to inductive risk. Thus, while ambitious variants of the democratic approach create more autonomy at the expense of practicality, the less ambitious variants are more practical, but allow for less autonomy.

What does this mean for APr? I would argue that if there were only one way to address prescriptiveness, this would undermine AIR's claim that *D* should sometimes involve extra-scientific utilities. However, while none of the above strategies is satisfying on its own, in conjunction they provide a feasible set of means to legitimize non-epistemic judgements in science. In some cases, it will be possible to circumvent the relevance condition by making uncontroversial value-judgements, or by using scenario sets that represent the spectrum of existing value commitments (Edenhofer & Kowarsch, 2015). In the remaining cases, there are ways to avoid the externality condition. We may here think of a multi-layer system (Steel, 2016a), where stakeholders contribute intensely to those studies that are closely entangled with extra-scientific decisions, e.g. advisory reports or technology assessments (Sclove, 2011; Garard & Kowarsch, 2017), but contribute to everyday science only on a higher level (e.g. via general guidelines, see Steel, 2016a). If additional judgements are needed in a concrete study, e.g. to interpret the general guidelines or to make choices that are not covered by the guidelines, scientists can

use the transparency approach to create some autonomy *qua* recipient. Therefore, while APr is right to emphasize the danger of prescriptiveness, this danger can be countered. Suitable measures against prescriptiveness exist, and as long as these are taken, APr does not refute AIR.

5.2 AWT's CHARGE OF WISHFUL THINKING

The second concern behind VFI, the argument from wishful thinking (AWT), holds that *D* is logically fallacious if it includes extra-scientific utilities. The main claim reads (see sect. 2.2):

AWT (2) If *D* is value-laden, *D* represents a *direct*, *non-vacuous* and *not semantically entailed* inference from an ought-claim to an is-claim (which violates no-is-from-ought).

Similar to APr, AWT hinges on three conditions. I restrict my discussion to these conditions, thus presuming that, if they are jointly fulfilled, *D* is indeed fallacious. Taking up meta-ethical work (Schurz, 1997; D'Arms & Jacobson, 2000; Pidgen, 2010; Pidgen, 2016), I understand AWT's conditions as follows: an ought-is inference is *direct* if the propositional source of the descriptive conclusion lies exclusively in a set of normative premises (e.g. "x should be the case, therefore x is the case"); an ought-is inference is *non-vacuous* if the descriptive conclusion is a substantial or non-arbitrary implication of a set of normative premises (an example of a vacuous inference is "x should be the case, therefore x should be the case or x is the case") (Prior, 1960; Pidgen, 2010); an ought-is inference is *not semantically entailed* if the descriptive conclusion is not hidden in the set of normative premises (an example of a semantically entailed inference is "x should be done, therefore x can be done") (Searle, 1964; Pidgen, 2016).

To see whether *D* fulfills these conditions, imagine that the agent concludes that it is justified to use a certain model, and that this conclusion is value-laden in the previously discussed sense ($T < 1$). The agent's set of premises would then comprise four types of elements:

D_1 A descriptive premise that specifies the probability *p* that using the model leads to truth.

D_2 A set of descriptive premises that specify the model's potential scientific and extra-scientific consequences, as well as these consequences' dependent probability *P*.

N_1 A set of normative premises that specify the utility *U* of the model's scientific consequences.

N_2 A set of normative premises that specify the utility *U* of the model's extra-scientific consequences.

It may be objected that the model choice cannot be subject to no-is-from-ought, as "it is justified to use the model" is not a descriptive conclusion. However, from the perspective of the "perfect scientist" this is just short for "before the background of the available evidence and the expected consequences, using the model promises more or at least equal benefits for science than not using the model". If we now accept AWT's premise that science should aspire truth (AWT-1), the conclusion commits the agent to the descriptive claim that the model helps to find scientific truths. Note that this is compatible with the idea that different versions of science may aspire different kinds of truth. In some contexts it may, e.g., be rational to prefer a less precise over a more precise model (Elliott & McKaughan, 2014). But this neither means that such a choice does not benefit science (Steel, 2016b), nor that the less precise model is not supposed to be truth-conducive. Rather, the model is supposed to help find exactly the kind of scientific truths that are deemed relevant in a given context.

So is the model choice fallacious in AWT's sense? As it turns out, two of AWT's conditions are fulfilled: the inference is non-vacuous because the descriptive proposition is not arbitrarily attached to the conclusion (as in "the model is ethically good, therefore the model is ethically good or the model is truth-conducive"); the inference is also not semantically entailed, i.e. not derived from an implicit descriptive content of a normative premises (as in "the model should be used, therefore the model can be used"). The third condition, *directness*, is more ambiguous. We may argue that directness is not given because the conclusion's descriptive content originates not from N_2 (the set of premises that represent the non-epistemic value-judgement), but from D_1 (the premise that characterizes the evidence). Douglas (2000, 2008, 2009) has argued into this direction. The "virtue of truth-seeking" (2008, p. 10), she holds, precludes non-epistemic values from acting "as reasons in themselves to accept a claim" (2009, p. 96); rather, their role is to "weigh the importance of uncertainty" (ibid.). On this account, the inference is not an instance of "the model is ethically good, therefore the model is truth-conducive" because the model's truth-conduciveness is inferred from a descriptive rather than from a

normative premise (Kevin Elliott has called this the "logical interpretation" of Douglas' approach, see 2013, p. 377).

I see two problems with this defense against AWT. First, it remains possible that non-epistemic values only "weigh the importance of uncertainty" (Douglas, 2009, p. 96) and yet dominate the evidence. This can occur when the expected extra-scientific utilities are distributed very unevenly, e.g. when the expected error damages are very high, while the expected truth benefits are very low (or vice versa). In such a scenario, utility distributions are conceivable where the agent *never* (or *always*) uses the model, irrespective of how well (or poorly) the model is supported by the evidence. Metaphorically speaking: if only the scale that weighs the evidence is sensitive or insensitive enough, any amount of evidence will be sufficiently "heavy" or "light" to justify a choice. Yet, if an evidential threshold is never (or always) met, D_1 is obviously irrelevant. As the only plausible source for the conclusion's descriptive content would then be N_2, AWT's directness condition can be fulfilled even if non-epistemic values only "weigh the importance of uncertainty" (ibid.). Secondly, Douglas' approach allows for *inverse preference orders*, i.e. scenarios where the agent prefers error over truth, and missed truth over averted error. This is because not all truths are extra-scientifically good, while not all errors are extra-scientifically bad. Placebo drugs, e.g., can have positive effects not although, but exactly because they are used on the basis of a false belief. However, if the agent prefers a convenient error over an inconvenient truth, she will use the model when p is low and dismiss it when p is high. Such a paradox notion of the evidence would undermine the claim that the conclusion stems from D_1. The possibility of inverse preference orders thus provides further support for AWT.

It could be objected that both issues, polarized utility distributions and inverse preference orders, are untypical. However, the claim is not that these issues occur often, but that, if they occur, they cannot be prevented by restricting non-epistemic values to determining evidential thresholds. Furthermore, the first issue can be relevant even if an evidential threshold is not *conceptually* impossible to meet or miss; it can suffice that a threshold is *practically* never met or missed in a given area of study. We may call such choices *material* rather than formal ought-is fallacies, as D_1 is still present in the set of premises, but practically irrelevant. One might then object that such choices can still be acceptable, as science has legitimate goals besides truth (Elliott, 2013; Elliott & McKaughan, 2014). As argued before, however, goals such as applicability or timeliness are not unrelated to truth, but qualify the kinds of truths that are aspired in a given context. Yet, I agree with the objection in one respect: some scientific choices may be ethically

impermissible, irrespective of how unlikely an error is. But this does not undermine my point that the mentioned issues are problematic. For one, this reasoning seems inapplicable to the other scenarios (e.g. truths that are so desirable that a choice is always made, or truths that are so undesirable that they are valued less than errors). For another, if a choice is morally impermissible, it should be seen as a boundary condition, similar to ethical norms in human trials. However, such norms are not subject to AWT or VFI in the first place (via VFI-R_2). Thus, while ethics may indeed sometimes trump truth-seeking, this does not help against AWT.

Another way to address AWT is to return to the previously discussed cases. I have argued that the "perfect scientist" should consider non-epistemic values only to resolve epistemic indifference (case A). Now, if "epistemic indifference" means that both options promise equal scientific benefits, and if "scientific benefit" means that an option helps to find scientific truths, then the conclusion's descriptive content is clearly not derived from N_2. As non-epistemic values only decide the choice between options whose truth-conduciveness has already been established, AWT's directness condition is not fulfilled in case A. Additionally, it seems implausible that scientific utilities could force evidentially unsupported choices. It is hard to see how an unattainable or unmissable evidential threshold could benefit science, as this would either add falsehoods to the body of scientific beliefs or make it impossible to find scientific truths. The same holds true for inverted preference orders. Even if some errors may be scientifically fruitful, the *deliberate* adoption of a false belief seems incompatible with the truth-seeking nature of science (note that we are not taking about simplifications or counterfactual assumptions here, as long as these are used to find scientific truths; inverted preferences, on the other hand, mean that a choice is made to purposely generate scientific falsehoods). As the directness condition is not fulfilled in case A, AWT does not succeed in this case.

So what about scenarios where the agent has a clear epistemic preference (case B)? I have argued that, at least for the "perfect scientist", VFI_{norm} remains valid in case B. However, this is because the "perfect scientist" cannot trade scientific for extra-scientific benefits, not because the use of non-epistemic values is *necessarily* fallacious in case B. I have already discussed that scientific and extra-scientific expected utilities may well pull into the same direction. Similar to case A, I would argue that such non-trade-off scenarios do not constitute ought-is fallacies, as non-epistemic values merely confirm an independent epistemic preference. AWT's directness conditions is hence not fulfilled in these scenarios. Nevertheless, the condition can indeed be fulfilled in the other scenario of case B, namely when scientific and extra-scientific utilities pull into opposite direc-

tions. If non-epistemic values are used in such trade-off scenarios, and if they change a decision from, say, using a model to not using a model, then the conclusion is clearly derived from N_2. Note again that such ethics-driven decisions may sometimes be acceptable in actual science. As said before, however, they should then be treated as ethical boundary conditions. Hence, while AWT would indeed succeed if ethics-driven decisions are interpreted as descriptive conclusions, AWT's charge of wishful thinking can be averted by seeing them as what they are: moral rather than genuinely scientific choices.

6. Conclusion: idealized *versus* actual science

Inductive risk is widely recognized as "[o]ne of the most important reasons for thinking that non-epistemic values can play a legitimate role in scientific reasoning" (Elliott & Steel, 2017, p. 6). A great part of AIR's appeal lies in its promise to refute VFI even under idealized assumptions. As Rudner has put it, AIR claims that VFI fails even for the "perfect scientist" or "the scientist *qua* scientist" (1953, p. 2). Not only is it "more surprising in the ideal setting that scientists must make value judgments" (Steele 2012, p. 895), such idealizations also capture the *in principle* nature of VFI (Weber, 1904/1949; Popper, 1976; Koertge, 2000; Ruphy, 2006; Kitcher, 2011; Reiss & Sprenger, 2020).

Taking up this challenge, I have proposed a Bayesian framework that accounts for subjective probabilities, outcomes other than error (Wilholt, 2009; Wilholt, 2013), and the difference between first- and second-order outcomes. The approach also gives us a clearer decision rule than more classic takes on AIR and eases the old problem (Kuhn, 1977) that epistemic values can stand in tension with each other. Finally, the trade-off parameter T represents the balancing problem of epistemic versus non-epistemic values in a more fine-grained way than, e.g., the notion of values "trumping" (Elliott & McKaughan, 2014) each other. Using the idealized setting as a testing ground, I have argued that AIR does not refute $\mathrm{VFI}_{\mathrm{desc}}$. Regarding $\mathrm{VFI}_{\mathrm{norm}}$, I have argued that AIR fails whenever the agent has a clear epistemic preference (case B), but succeeds whenever the expected scientific utilities of the decision options converge (case A). I have argued that the notion of utility convergence goes beyond common versions of the "tie breaker" thesis (Intemann, 2005; Brown, 2013; Magnus, 2018), and that utility convergence is more typical than it may seem at first. Hence, while AIR's refutation of VFI is not *complete*, it still represents a powerful critique of val-

ue-freedom. This is further supported by the fact that two of the main concerns behind VFI, APr and AWT, can be countered by avoiding these arguments' conditions. It may be argued that my decision setting is unrealistic and, hence, practically irrelevant. I disagree. In fact, many of its aspects are surprisingly realistic. First, inductive risk involves more than, e.g., the classic distinction between consumer versus producer risks (e.g. Carrier, 2011; Biddle & Leuschner, 2015). In reality, any given scientific choice can have many consequences; these will have different (dependent) probabilities and will occur not only in case of error (e.g. missing a truth is not an error, but can be ethically relevant). Second, the classic inductive risk heuristic "worse consequences = higher evidential thresholds" is not more, but less practical than the Bayesian reconstruction, as it neither contains a point of reference nor an idea of how exactly these thresholds should be determined. Third, it is quite realistic to interpret epistemic indifference as utility convergence, as a decision's scientific worth will *practically* also depend on the likelihood and the desirability of its scientific consequences. Fourth, it is very realistic to assume that probabilities are typically subjective in a Bayesian sense; also, probabilities will often be imprecise (Parker, 2014), which makes epistemic indifference, and hence my arguments regarding case A, highly relevant in actual science. This is also why the objection that scientific utilities may be unclear or disputed in practice does not speak against my reconstruction; this simply means that utilities can come as intervals as well, which in turn makes case A even more relevant. Finally, the balancing problem represented by T represents a quite practical issue, as scientists cannot make inductive risk decisions if the relative weight of the epistemic and non-epistemic values remains unclear.

Thinking this a bit further, I would argue that the idealized setting outlines something that may be called *epistemic legitimacy* – a set of rules that should govern the use of non-epistemic values in actual science. Similar to Steel (2010)[13], I contend that it should be the standard approach for scientists to use non-epistemic values only in cases of epistemic indifference. As we can see from the discussion of AWT, scientists should always favor truth over error and averted error over missed truth, irrespective of how desirable a decision's second-order outcomes are. Furthermore, scientists must carefully avoid illegitimate prescription, typically by applying a combination of the approaches presented in the discussion of APr. While I believe that scientists should normally not use non-epistemic values if they have a clear epistemic preference, I also concede that ethical concerns may *sometimes* outweigh scientific considerations (e.g. when the extra-scien-

13 Note that Steel (2010) interprets epistemic indifference in a different way, namely as a balance between epistemic values. As said before, however, epistemic indifference is better captured as a convergence of expected scientific utilities.

tific consequences are both very bad and very likely). In such cases, scientists should scrutinize whether there is a real trade-off, i.e. whether epistemic and non-epistemic values actually pull into different directions. If there is a real trade-off, scientists should use non-epistemic values only if the expected extra-scientific benefit is significantly higher than the expected scientific loss. I would also argue that the relative weight that non-epistemic values can have in a trade-off scenario, i.e. *T*, should not be determined by individual scientists or research groups, but by codified guidelines (issued by, e.g., national academies or research associations). Most importantly, if scientists make decisions against an epistemic preference, they must communicate this as an *ethical* rather than a scientific choice.

Acknowledgements

I would like to thank Torsten Wilholt, Dietmar Hübner, Roel Visser, Hermann Held, and the members of the research training group GRK 2073 "Integrating Ethics and Epistemology of Scientific Research" for their support and helpful comments. I particularly wish to thank Benjamin Blanz for his generous support on the decision-theoretical part. This paper also benefited from discussions with members of the Consortium for Science, Policy and Outcomes at Arizona State University and the Mercator Research Institute on Global Commons and Climate Change, particularly Martin Kowarsch. Funding was provided by the German Research Foundation (DFG) (254954344/GRK2073) and the German Academic Exchange Service (DAAD).

References

Beck, S. (2011). Moving beyond the linear model of expertise? IPCC and the test of adaptation. *Regional Environmental Change*, 11(2), 297–306. https://doi.org/10.1007/s10113-010-0136-2

Beeson, M. (2010). The coming of environmental authoritarianism. *Environmental Politics*, 19(2), 276–294. https://doi.org/10.1080/09644010903576918

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, 3(2), 207–220. https://doi.org/10.1007/s13194-012-0062-x

Biddle, J. (2013). State of the field: Transient underdetermination and values in science. *Studies in History and Philosophy of Science*, 44(1), 124–133. https://doi.org/10.1016/j.shpsa.2012.09.003

Biddle, J., & Kukla, R. (2017). The Geography of Epistemic Risk. In K. Elliott, & T. Richards (Eds.), *Exploring Inductive Risk: Case Studies of Values in Science* (pp. 215–237). Oxford University Press.

Biddle, J., & Leuschner, A. (2015). Climate Skepticism and the manufacture of doubt: can dissent in science be epistemically detrimental? *European Journal for Philosophy of Science*, 5, 261–278.

Biddle, J., & Winsberg, E. (2010). Value Judgements and the Estimation of Uncertainty in Climate Modelling. In P.D. Magnus, & J. Busch (Eds.), *New Waves in Philosophy of Science* (pp. 172–197). Palgrave Macmillan.

Bright, L. (2018). Du Bois' democratic defence of the value free ideal. *Synthese*, 95(5), 2227–2245.

Brown, M. B. (2009). *Science in democracy. Expertise, institutions, and representation*. MIT Press.

Brown, M. J. (2013). Values in Science beyond Underdetermination and Inductive Risk. *Philosophy of Science*, 80, 829–839. https://doi.org/10.1086/673720

Bueter, A. (2015). The irreducibility of value-freedom to theory assessment. *Studies in History and Philosophy of Science*, 49, 18–26. https://doi.org/10.1016/j.shpsa.2014.10.006

Carrier, M. (2011), Knowledge, Politics, and Commerce: Science Under the Pressure of Practice. In M. Carrier, & A. Nordmann (Eds.), *Science in the Context of Application* (pp. 11–30). Springer.

Churchman, C. W. (1948). *Theory of Experimental Inference*. Macmillan.

D'Arms, J., & Jacobson, D. (2000). The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophy and Phenomenological Research*, 61(1), 65–90.

Davies, C., Wetherell, M., Barnett, E., & Seymour-Smith, S. (2005). *Opening The Box. Evaluating the Citizens Council of NICE*. The Open University.

De Melo-Martín, I., & Intemann, K. (2016). The Risk of Using Inductive Risk to Challenge the Value-Free Ideal. *Philosophy of Science*, 83, 500–520. https://doi.org/10.1086/687259

Dorato, M. (2004). Epistemic and Nonepistemic Values in Science. In P. Machamer, & G. Wolters (Eds.), *Science, Values, and Objectivity* (pp. 52–77). University of Pittsburgh Press.

Douglas, H. E. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, 67(4), 559–579. https://doi.org/10.1086/392855

Douglas, H. (2005). Inserting the Public Into Science. In S. Maasen, & P. Weingart (Eds.), *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making* (pp. 153–169). Springer.

Douglas, H. (2008). The Role of Values in Expert Reasoning. *Public Affairs Quarterly*, 22(1), 1–18.

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Douglas, H. (2015), Values in Science. In P. Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science* (pp. 609–630). Oxford University Press.

Douglas, H. (2017). Why Inductive Risk Requires Values in Science. In K. Elliott, & D. Steel (Eds.), *Current Controversies in Values and Science* (pp. 81–93). Routledge.

Du Bois, W. E. B. (1935). *Black reconstruction in America*. The Free Press.

Dupré, J. (2007). Fact and Value. In H. Kincaid, J. Dupré, & A. Wylie (Eds.), *Value-Free Science? Ideals and Illusions* (pp. 27–41). Oxford University Press.

Edenhofer, O., & Kowarsch, M. (2015). Cartography of pathways: A new model for environmental policy assessments. *Environmental Science & Policy*, 51, 56–64. https://doi.org/10.1016/j.envsci.2015.03.017

Elliott, K. C. (2011). *Is a little pollution good for you? Incorporating societal values in environmental research*. Oxford University Press.

Elliott, K. C. (2013). Douglas on values: From indirect roles to multiple goals. *Studies in History and Philosophy of Science,* 44, 375–383. https://doi.org/10.1016/j.shpsa.2013.06.003

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic Values and the Multiple Goals of Science. *Philosophy of Science*, 81(1), 1–21. https://doi.org/10.1086/674345

Elliott, K. C., & Steel, D. (2017). Introduction: Values and Science: Current Controversies. In K. Elliott, & D. Steel (Eds.), *Current Controversies in Values and Science* (pp. 1–11). Routledge.

Garard, J., & Kowarsch, M. (2017). Objectives for Stakeholder Engagement in Global Environmental Assessments. *Sustainability*, 9, 1571.

Grundmann, R., & Rödder, S. (2019). Sociological Perspectives on Earth System Modeling. *Journal of Advances in Modeling Earth Systems*, 11(12), 3878–3892. https://doi.org/10.1029/ 2019MS001687

Haack, S. (2003). Knowledge and Propaganda. Reflections of an Old Feminist. In C.

Pinnick, N. Koertge, & R. Almeder (Eds.), *Scrutinizing Feminist Epistemology. An Examination of Gender in Science* (pp. 7–19). Rutgers University Press.

Habermas, J. (1979). What is Universal Pragmatics? In *Communication and the Evolution of Society* (pp. 1–68). Beacon Press.

Habermas, J. (1998). Some further clarifications of the concept of communicative rationality. In *On the pragmatics of communication* (pp. 307–342). Polity Press.

Hassani, H., Beneki, C., Silva, E. S., Vandeput, N., & Madsen, D. Ø. (2021). The science of statistics versus data science: What is the future?, *Technological Forecasting and Social Change*, 173, 121111.

Held, H. (2011). Dealing with Uncertainty – From Climate Research to Integrated Assessment of Policy Options. In G. Gramelsberger, & J. Feichter (Eds.), *Climate Change and Policy. The Calculability of Climate Change and the Challenge of Uncertainty* (pp. 113–126). Springer.

Hempel, C. (1965). Science and Human Values. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science* (pp. 81–96). The Free Press.

Hempel, C. (1981). Turns in the evolution of the problem of induction. *Syntheses* 46, 389–404.

Holman, B., & Wilholt, T. (2022). The New Demarcation Problem. *Studies in History and Philosophy of Science*, 91, 211–220. https://doi.org/10.1016/j.shpsa.2021.11.011

Hoyningen-Huene, P. (2006). Context of Discovery versus Context of Justification and Thomas Kuhn. In J. Schickore, & F. Steinle (Eds.), *Revisiting Discovery and Justification* (pp. 119–131). Springer.

Hume, D. (2009). *A treatise of human nature. Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. Oxford University Press. (First publ. 1739)

Intemann, K. (2005). Feminism, underdetermination, and values in science. *Philosophy of science*, 72(5), 1001–1012. https://doi.org/10.1086/508956

Intemann, K. (2015). Distinguishing between legitimate and illegitimate values in climate modeling. *European Journal for Philosophy of Science*, 5, 217–232. https://doi.org/10.1007/s13194-014-0105-6

James, W. (1912). The Will to Believe. In *The Will To Believe and other Essays in Popular Philosophy* (pp. 1–31). Longmans, Green & Co.

Jasanoff, S., & Wynne, B. (1998). Science and Decisionmaking. In S. Rayner, & E. Malone (Eds.), *Human Choice and Climate Change. Vol 1: The Societal Framework* (pp. 1–87). Battelle Press.

John, S. (2015). Inductive risk and the contexts of communication. *Synthese*, 192(1), 79–96. https://doi.org/10.1007/s11229-014-0554-7

John, S. (2019). Science, truth and dictatorship: Wishful thinking or wishful speaking?. *Studies in History and Philosophy of Science*, 78, 64–72. https://doi.org/10.1016/j.shpsa.2018.12.003

Jones, O. (1999). Sex, Culture, and the Biology of Rape: Toward Explanation and Prevention. *California Law Review*, 87(4), 827–941.

Kitcher, P. (2001). *Science, truth, and democracy*. Oxford University Press.

Kitcher, P. (2011). *Science in a democratic society*. Prometheus Books.

Koertge, N. (2000). Science, Values, and the Value of Science. *Philosophy of Science*, 67(S3), S45–S57. https://doi.org/10.1086/392808

Kourany, J. A. (2003). A Philosophy of Science for the Twenty-First Century. *Philosophy of Science*, 70(1), 1–14. https://doi.org/10.1086/367864

Kowarsch, M., Garard, J., Riousset, P., Lenzi, D., Dorsch, M., Knopf, B., Harrs, J., & Edenhofer, O. (2016). Scientific assessments to facilitate deliberative policy learning. *Palgrave Communications*, 2(1), 1–20. https://doi.org/10.1057/palcomms.2016.92

Kuhn, T. (1977). Objectivity, Value Judgement, and Theory Choice. In *The Essential Tension. Selected Studies in Scientific Tradition and Change* (pp. 320–339). University of Chicago Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lacey, H. (1999). *Is science value free? Values and scientific understanding*. Routledge.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. (2008). Values, Heuristics, and the Politics of Knowledge. In M. Carrier, D. Howard, & J. Kourany (Eds.), *The Challenge of the Social and the Pressure of the Practice* (pp. 68–86). University of Pittsburgh Press.

Lovelock, J. (2010, March 29). *James Lovelock on the value of sceptics and why Copenhagen was doomed* (Interview). The Guardian. https://www.theguardian.com/environment/blog/2010/mar/ 29/ james-lovelock

Machamer, P., & Douglas, H. (1999). Cognitive and social values. *Science & Education*, 8(1), 45–54.

McMullin, E. (1982). Values In Science. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. Two: Symposia and Invited Papers, 3–28. https://doi.org/10.1086/psaprocbienmeetp.1982.2.192409

Magnus, P. D. (2018). Science, Values, and the Priority of Evidence. *Logos & Episteme*, 9(4), 413–431. https://doi.org/10.5840/logos-episteme20189433

Parker, W. (2014). Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science*, 46, 24–30.

Pidgen, C. (2010). On the Triviality of Hume's Law: a Reply to Gerhard Schurz. In C. Pidgen (Ed.), *Hume on Is and Ought* (pp. 212–236). Palgrave Macmillan.

Pidgen, C. (2016). Hume on Is and Ought. In P. Russel (Ed.), *The Oxford Handbook of Hume* (pp. 401–415). Oxford University Press.

Popper, K. (1976). The Logic of the Social Sciences. In T. Adorno, H. Albert, R. Dahrendorf, J. Habermas, H. Pilot, & K. Popper (Eds.), *The Positivist Dispute in German Sociology* (pp. 87–104). Heinemann.

Prior, A. (1960). The autonomy of ethics. *Australasian Journal of Philosophy*, 38(3), 199–206. https://doi.org/10.1080/00048406085200221

Proctor, R. (1991). *Value-free science? Purity and power in modern knowledge*. Harvard University Press.

Putnam, H. (2002). *The Collapse of the Fact/value Dichotomy and Other Essays*. Harvard University Press.

Reichenbach, H. (1961). *Experience and Prediction*. University of Chicago Press. (First publ. 1938)

Reiss, J., & Sprenger, J. (2020). Scientific Objectivity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). https://plato.stanford.edu/entries/scientific-objectivity/

Rudner, R. (1953). The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science*, 20(1), 1–6. https://doi.org/10.1086/287231

Ruphy, S. (2006). Empiricism all the way down: a defense of the value-neutrality of science in response to Helen Longino's contextual empiricism. *Perspectives on Science*, 14(2), 189–214. https://doi.org/10.1162/posc.2006.14.2.189

Schuessler (2019). *The Debate on Probable Opinions in the Scholastic Tradition*. Brill.

Schurz, G. (1997). *The Is-Ought-Problem*. Springer.

Sclove, R. (2010). Reinventing Technology Assessment. *Issues in Science and Technology*, 27(1), 34–38.

Searle, J. R. (1964). How to Derive "Ought" From "Is". *The Philosophical Review*, 73(1), 43–58.

Shearman, D. J. C., & Smith, J. W. (2007). *The climate change challenge and the failure of democracy*. Praeger Publishers.

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, 77, 14–34. https://doi.org/10.1086/650206

Steel, D. (2016a). Climate Change and Second-Order Uncertainty: Defending a Generalized, Normative, and Structural Argument from Inductive Risk. *Perspectives on Science*, 24(6), 696–721. https://doi.org/10.1162/POSC_a_00229

Steel, D. (2016b). Accepting an Epistemically Inferior Alternative? A Comment on Elliott and McKaughan. *Philosophy of Science*, 83, 606–612. https://doi.org/10.1086/687264

Steele, K. (2012). The Scientist qua Policy Advisor Makes Value Judgments. *Philosophy of Science*, 79(5), 893–904. https://doi.org/10.1086/667842

Tomblin, D., Pirtle, Z., Farooque, M., Sittenfeld, D., Mahoney, E., Worthington, R., Gano, G., Gates, M., Bennett, I., Kessler, J., Kaminski, A., Lloyd, J., & Guston, D. (2017). Integrating Public Deliberation into Engineering Systems: Participatory Technology Assessment of NASA's Asteroid Redirect Mission. *Astropolitics*, 15(2), 141–166. https://doi.org/10.1080/14777622.2017.1340823

Weber, M. (1949). "Objectivity" in Social Science and Social Policy. In *On the Methodology of the Social Sciences* (pp. 50–112). The Free Press. (First publ. 1904)

Weber, M. (1958). Science as a Vocation. *Daedalus*, 87(1), 111–134. (First publ. 1919)

Weingart, P. (1999). Scientific expertise and political accountability: paradoxes of science in politics. *Science and Public Policy*, 26(3), 151–161. https://doi.org/10.3152/147154399781782437

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science*, 40(1), 92–101. https://doi.org/10.1016/j.shpsa.2008.12.005

Wilholt, T. (2013). Epistemic Trust in Science. *British Journal for the Philosophy of Science*, 64, 233–253. https://doi.org//10.1093/bjps/axs007

Wilholt, T. (2016). Collaborative Research, Scientific Communities, and the Social Diffusion of Trustworthiness. M. Brady, & M. Fricker (Eds.), *The Epistemic Life of Groups Essays in the Epistemology of Collectives* (pp. 218–249). Oxford University Press.

Winsberg, E. (2012). Values and Uncertainties in the Prediction of Global Climate Models. *Kennedy Institute of Ethics Journal*, 22(2), 111–137. https://doi.org/10.1353/ken.2012.0008

IV. CITIZEN PARTICIPATION IN KITCHER'S WELL-ORDERED SCIENCE: WHAT KIND OF IDEAL IS DELIBERATION?

ZUSAMMENFASSUNG Im dritten Teil habe ich einige Strategien zur Umgehung des Präskriptionsarguments aufgezeigt. Die Einbeziehung außerwissenschaftlicher Stakeholder spielt hierbei eine zentrale Rolle. Im vierten Teil diskutiere ich einen der wichtigsten philosophischen Beiträge zum Thema der Stakeholderintegration: den Ansatz Philip Kitchers. Nach Kitcher soll das Wertfreiheitsideal durch ein *deliberatives Ideal* ersetzt werden, welches die Akzeptabilität einer Forschungsentscheidung an die (tatsächliche oder hypothetische) Zustimmung aller derjenigen bindet, die von dieser Entscheidung betroffen sind oder sein können. Hierfür muss eine Reihe anspruchsvoller Bedingungen wie universelle Inklusion, vollständige Gleichheit oder perfekte Unparteilichkeit erfüllt sein. Wie Kitcher jedoch selbst betont, sind diese Bedingungen unerfüllbar. Sein Ansatz setzt sich daher der *realistischen Kritik* aus, dass das deliberative Ideal (i) nicht verpflichtend, (ii) anfällig für nicht-intendierte Folgen und (iii) praktisch irrelevant ist. Dies ist problematisch, da Kitcher ebendiese Kritik gegen das Wertfreiheitsideal verwendet. Ich diskutiere dieses Problem anhand einer Unterscheidung zwischen vier Lesarten des Begriffs „Ideal". Ich zeige, dass die plausibelste Erwiderung auf die realistische Kritik in dem besteht, was ich die *rekonstruktive Lesart* nenne.

1. Introduction: Philip Kitcher and the axiological turn

In recent years, philosophy of science has experienced what we may call an *axiological turn*[1]: an increasing interest in, and acceptance of, extra-scientific values in science (Douglas, 2000; Longino, 2004; Kourany, 2008; Willholt, 2009; Elliott, 2011; Biddle, 2013). Many philosophers today agree that "[w]e should stop asking whether social values play a role in science and instead ask which values and whose values play a role and how" (Longino, 2004, p. 127). However, while "the positive project of determining more precisely the roles of values in science" (Brown, 2013, p. 835) has doubtlessly gained momentum, the specifics of such a value-embracing "philosophy of science for the twenty-first century" (Kourany, 2003) are yet to be determined (Holman & Wilholt, 2022).

One major contribution to this ongoing debate is Philip Kitcher's book *Science in a Democratic Society* (2011). Kitcher argues that the traditional scientific ideal, the ideal of value-freedom, should be replaced by what he calls "well ordered science". The centerpiece of Kitcher's well-ordered science is a concept of "ideal deliberation under conditions of mutual engagement" (ibid., p. 51)[2]. This concept, which I refer to as the *deliberative ideal* or simply Kitcher's ideal, holds that value-laden scientific choices are legitimate if and only if they emerge (or would hypothetically emerge) from a discussion among everyone who is potentially affected by these choices. However, the requirements of ideal deliberation, such as universal inclusion or perfect impartiality, are rarely

1 I borrow this term from Peter Kroes and Anthonie Meijers (2016), who describe a similar turn in the philosophy of technology. By "axiological turn", I mean those positions in the philosophy of science that argue that non-epistemic values either *de facto* play an inevitable role in scientific activities such as hypothesis assessment, data analysis or model choice, or that some non-epistemic values should *ideally* play such a role. This comprises a diversity of positions (e.g. Harding, 1995; Longino, 2002; Kourany, 2003; Douglas, 2009; Wilholt, 2009; Elliott, 2011; Kitcher, 2011). The term should not be confused with a more specific differentiation that Bennett Holman and Torsten Wilholt (2022) recently made, where "axiological strategy" describes a particular way of discerning acceptable from unacceptable values in science.

2 Kitcher calls both well-ordered science and deliberation an ideal. Strictly speaking, well-ordered science is the umbrella term that includes a range of considerations, whereas deliberation is a subordinate term that specifies well-ordered science in the domain of decision-making procedures. In this paper, I will focus exclusively on the ideal of deliberation, as the question of how an ideal decision-making procedure would look like is one of the most important questions, or even *the* most important question, in Kitcher's well-ordered science.

(if ever) met in reality. This raises the question whether Kitcher's ideal is too unrealistic to guide actual science. Points of critique may be that the deliberative ideal is (i) an unattainable, hence normatively non-binding goal; (ii) an overambitious project that may cause more problems than it solves; or (iii) a naïve dream without practical relevance. I call this the *realist challenge*.

In this paper, I scrutinize the realist challenge by distinguishing four notions of "ideal": the *blueprint*, the *compass*, the *yardstick* and the *reconstructive* notion. I argue that the reconstructive notion of ideal deliberation offers the most promising reply to the realist challenge. I also show that the realist challenge is not some far-fetched consideration, but is present in many popular critiques of the value-free ideal, including Kitcher's own. In order to avoid inconsistency, Kitcher's alternative should therefore employ a notion of "ideal" that is less vulnerable to these realist arguments. In addition to defending a reconstructive notion of the deliberative ideal, my aim is to emphasize the conceptual differences between different types of ideals. Philosophers of the axiological turn have long focused on whether or not the value-free ideal is sound, and how it may be replaced by an alternative ideal. Apart from brief remarks, however, it has not always been clear what is meant by "ideal" in the first place. This is unfortunate, as different understandings of the term result in different options for defending and criticizing an ideal. Scrutinizing these understandings is therefore crucial, not only for supporters of Kitcher's ideal, but for the axiological turn more generally.

2. The direction of fit: should ideals be realistic?

Before I discuss Kitcher's ideal of deliberation, let me frame the question more generally. Why would we want ideals to be realistic in the first place? Is it not in the nature of ideals to describe a state of perfection that we can only approximate, but never fully achieve? In this vein, Kant held that an ideal "exists merely in thoughts" (CPR, A 569/B 597); it "serves as the original image" or *Urbild*, "with which we can compare ourselves, judging ourselves and thereby improving ourselves, even though we can never reach the standard" (ibid.). As it turns out, however, ideals can also be understood in a more realist manner. Take the ideal that has been at the heart of many critical debates in the wake of the axiological turn: the ideal of value-free science. While it is true that some authors

have treated value-freedom as "an ideal, which though we cannot perfectly adhere to it, should nevertheless be held as an aspiration" (Bright, 2018, p. 2244; see also Weber, 1904/1949; Popper, 1976; McMullin, 1982), others have assumed a close link between value-freedom as an ideal and as an actual possibility (Rudner, 1953; Biddle, 2013; Reiss & Sprenger, 2020). Justin Biddle, for instance, has argued that the value-free ideal "states not only that we should strive to minimize the influence of contextual factors, but also that the proper application of scientific methods will always result in our ability to screen them out entirely" (2013, p. 131). We will see later that Kitcher also treats the issue of realism, and the lack thereof, as a major issue of the value-free ideal. It thus seems that the notion of ideals as essentially unrealistic ambitions is not so innocent after all – and that, consequently, there may be reasons why our ideals should be somewhat realistic.

In order to get a more systematic grip on this issue, I would like to borrow a term that goes back to Elisabeth Anscombe (1957) and that has later been popularized by John Searle (1983): the direction of fit. It describes the idea that different types of attitudes refer to the world in different ways. Factual beliefs "aim to have their content fit the world" (Gregory, 2012, p. 604), such that "when the content of a belief fails to match up to how reality is, it is up to the belief to change" (ibid.); in contrast, "[d]esires, and other desire-like attitudes, aim to have the world fit their content", such that "when the content of a desire fails to match up to how reality is, it is up to the world to change" (ibid.). As ideals are desire-like attitudes, is seems plausible that holding an ideal means to aim to adapt the world to the ideal rather than the ideal to the world. This is intuitive for ideals such as justice, prosperity, integrity and many others. We can thus assume that the relation between ideals and the world is organized along what we may call the *standard direction of fit*, such that the world is supposed to fit the ideal, while the ideal need not necessarily fit the world.

As said before, however, many scholars believe that considerations of realism are still relevant to ideals. In fact, it has often been argued that unrealistic goals are "akin to 'magical thinking' or 'science fiction'" (Guillemot, 2017, p. 49), and that a "mismatch between [normative] theory and practice may not simply mean that practice has not lived up to the theoretical ideals but perhaps that the theory demands too much of the real world" (Lövbrand et al., 2010, p. 14). If this is so, it will sometimes be necessary to adapt a given ideal to reality – to "reverse" the direction of fit, as it were[3]. I see three

3 To avoid misunderstandings, note that the idea of a "reversed direction of fit" is not meant to imply that an ideal is somehow supposed to represent reality in the same way as a factual belief would. The term "reverse"

major reasons why such a realist modification may be necessary in a given case, each of which is relevant to Kitcher's deliberative ideal and the axiological turn more generally:

i. In what we may call the *argument from non-bindingness*, we could maintain that the normative content of an ideal – the ideal's *ought* – refers to a specific state of the world rather than one that merely approximates it. If we cannot realize that specific state, we could conclude via ought-implies-can that such an ideal is not normatively binding. Biddle (2013) has used exactly this argument against the ideal of value-free science: "to claim that we ought to strive to achieve epistemic purity is to imply that we can achieve epistemic purity; if epistemic purity is impossible or unreasonable to achieve, however, then we should not maintain that we ought to strive to achieve it" (2013, p. 131). Biddle concedes that scientists may still approximate value-freedom; yet, he argues that this misses the normative content of this particular ideal, which essentially demands value-freedom rather than value-minimization (ibid.).
ii. A reasoning that we may call the *argument from unintended effects* holds that attempts to achieve the unachievable often lead to bad outcomes. This is acknowledged even by defenders of ambitious ideals. In the context of political ideals, e.g., David Estlund concedes that "[a]ctions in pursuit of what will never be achieved can be wasteful or even disastrous. A theory that counsels action in pursuit of high standards that are not sufficiently likely to be achieved, where the costs of failing are very high, often deserves to be chastised as utopian" (2014, p. 120). A similar reasoning has been used by critics of scientific value-freedom (e.g. Longino, 2002) who maintain that "the ideal can have the dangerous consequence of masking the influence of contextual factors" (Biddle, 2013, p. 131). More generally speaking, the argument claims that unrealistic ideals tend to create more problems than they solve.
iii. A third consideration, call it the *argument from irrelevance*, views ideals as means of social coordination, the function of which is to bring about ethically desired changes in our behavior. The argument claims that unrealistic ideals perform poorly in this function, as people either ignore these ideals or use them only to keep up the appearance. In a political context, Oliver Geden (2015) has used such an argument against what he calls "targetism". He claims that actual policy is not driven by moral ambitions, but by strategic considerations and by "the manifold limitations for 'optimal' policymaking" (Geden & Beck, 2014, p. 748) such as power or feasibility. Unrealistic ideals may also be irrelevant in other ways, namely when people think that

rather describes a process in which we lower our normative expectations, thus fitting a given ideal to the world. Having done so, our modified ideal will again be subject to the standard direction of fit.

> they comply with the ideal, while in reality they do not (because they cannot). This reasoning is implicit in critiques of scientific value-freedom that claim that "science, even science done well, is not value-free, and so the ideal is irrelevant" (Douglas, 2015, p. 5). In both cases, the claim is that an ideal is irrelevant if it does not impact our actual practices.

These arguments make up what I call the *realist challenge*. Note that, while all three of these claims address the issue of attainability, only the argument from non-bindingness targets this issue directly (i.e. unattainability is a reason in itself to reject an ideal). The other two arguments are more indirect, as they focus on the consequences of unattainable ideals – both the lack of desired consequences and the occurrence of undesired ones – rather than the unattainability itself. Note furthermore that there may well be other arguments against unrealistic ideals (for a related debate in political philosophy, see e.g. Simmons, 2010; Hamlin & Stemplowska, 2012; Valentini, 2012), as well as other versions of the above arguments. As we shall see shortly, however, the mentioned three arguments are crucial for both Kitcher's approach and the wider context of the axiological turn. In order to see this, we need to consider Kitcher's criticism of value-free science.

3. Science, democracy, and the failure of the value-free ideal

In *Science in a Democratic Society* (2011), Kitcher aims to determine the proper relation between science and society. For this, he distinguishes four types of scientific activities (ibid., p. 91): the specification of research agendas ("investigation"); the determination of scientific standards and procedures ("submission"); the acceptance of research findings ("certification"); and the dissemination and application of research findings ("transmission"). These activities constitute the "system of public knowledge" (ibid., p. 89). As not all citizens can be professional researchers, such a system entails a "division of epistemic labor" (ibid., p. 20) that regulates scientists' and citizens' roles in each of the four activities. Kitcher's question, then, is which division of labor best fits into the normative architecture of a democratic society.

However, there is a traditional answer to this question, and it is crucial to see why Kitcher rejects it. In the traditional view, the democratic citizenry has a say in what Kitcher

calls transmission and, perhaps to a lower degree, investigation. But when it comes to Kitcher's submission and certification stage, i.e. to deciding how science is to be conducted and what counts as a valid result, scientists enjoy full authority (Douglas, 2009, ch. 3). The traditional or "decisionist" model (Habermas, 1970; Weingart, 1999; Millstone, 2005) thus embraces a division of labor where those parts of science that the model assumes to be value-laden are (co-)determined by citizens and their representatives, whereas those parts that the model assumes to be value-free are controlled by scientists. However, this seems to presuppose that the latter parts do not involve the very kind of value-judgements that the model reserves for the democratic discourse. The traditional model thus rests on the *ideal of value-freedom*: the normative expectation that scientists do not allow extra-scientific values "at the heart of science" (Douglas, 2009, p. 15), i.e. in submission and certification.

As most philosophers of the axiological turn, however, Kitcher maintains that the value-free ideal, and hence the traditional division of labor, have ultimately failed. He draws on two classic arguments to show this. The first, known as the argument from inductive risk (Rudner, 1953), holds that research is a sequence of decision points, each of which forces the scientist to judge "whether what has been done so far is enough to warrant taking the next step" (Kitcher, 2011, p. 35). As empirical claims are uncertain, such decisions entail error risks. Whenever the consequences of error reach beyond science, the argument holds, scientific choices presuppose value-judgements about whether these consequences are acceptable (Douglas, 2000; Wilholt, 2009; Winsberg, 2012)[4]. The second argument claims that hypotheses are tested against the background of what is considered worth of investigation (Bueter, 2015). Kitcher here holds that judgments in the submission and certification stages are entangled with judgments in the investigation stage, which in turn depend on broader judgements on what a society deems valuable. He calls this "broad", "cognitive" and "probative" schemes of values (2011, p. 37) and illustrates their entanglement with paradigm changes as described by Kuhn (1962). Given this entanglement, Kitcher argues that scientific choices cannot be made independently from extra-scientific values.

4 The argument from inductive risk has been spelled out in more than one way. In particular, there is not only the *descriptive* version outlined above (which holds that value-freedom is impossible), but also a *normative* version (which holds that value-freedom is undesirable, irrespective of whether it is possible) (Betz, 2013; Steel, 2016). Similar to Betz and Steel, I have argued on a different occasion that the normative version is more promising than the descriptive version (Dressel, 2022). In our context, however, the descriptive version is more relevant as it is the one that Kitcher draws on, and because a realist critique always starts from descriptive rather than normative grounds.

In this paper, I will not take a stance on whether these arguments are valid. What I do want to maintain, however, is that even if they are, they do not *automatically* imply that the value-free ideal fails. This is where the three arguments of the realist challenge come in. As it turns out, two of these reasons are explicitly mentioned by Kitcher, while one is implicit in his ethical theory:

i. The argument from non-bindingness is present in one passage where Kitcher argues that "it is not just that many individual scientists do not live up to the standard of value-freedom, but that they cannot do so–and therefore it is not the case that they should do so" (2011, p. 31). Similar to Biddle (2013), Kitcher here argues that scientists cannot be held accountable for not achieving an unachievable standard. Later, however, he concedes that "scientists might try to minimize the value-judgments they make" (2011, p. 39). He even says that value-freedom may be seen "as a standard we might do well to approximate, when and to the extent we can" (ibid., p. 40). Thus, if it were only for the argument from non-bindingness, the value-free ideal may still have some place in Kitcher's well-ordered science.
ii. Yet, there is also the argument from unintended effects. This argument is a main driver in Kitcher's discussion of scientific authority, and perhaps the most important reason why he rejects the value-free ideal. Kitcher argues that citizens have been told that science is value-free, but that "the past decades have presented citizens with a sequence of examples in which dueling scientists could be viewed as importing values into […] their inquiries" (2011, p. 30). Today, Kitcher holds, many citizens are disappointed because "scientific practice typically does not live up to the standard of disinterested inquiry" (ibid.). The concerning effect is that citizens start to distrust science. However, this is only because scientists are measured against an unattainable standard: "The deepest source of the current erosion of scientific authority consists in insisting on the value-freedom of Genuine Science" (ibid., p. 40). Kitcher thus argues that upholding the value-free ideal, while at the same time being unable to realize it, has the unintended effect of undermining citizens' trust in science.
iii. Finally, there is the argument from irrelevance. Although Kitcher does not explicitly use this argument against value-freedom, it is well in line with his general ethical theory. Ethics, Kitcher holds, is a "social technology" (ibid., p. 47) to coordinate human activities, which in turn creates evolutionary advantages (ibid., p. 41-49). Thinking this a bit further, we may assume that an ideal is an effective "technology" if individuals actually believe in it, and that the strength of this conviction is, at least in part, a function of the ideal's perceived attainability. Consequently, we may conjecture that the value-free ideal has a weaker effect on those who perceive it as "a figment of

philosophical imagination" (Hedgecoe, 2004, p. 131). This is supported by studies that find that scientists who believe that value-freedom is unattainable also tend to reject value-freedom as an ideal (Reiners et al., 2013; Van der Hel, 2018). Some studies find this to be a minority (Steel et al., 2004), others even argue that "the value-free ideal is not a dominant perspective among scientists" (Steel et al., 2018, p. 639). But irrespective of how large this group actually is, these findings provide a *prima facie* reason for arguing that the value-free ideal would make for a more effective "social technology" if it were perceived as more realistic.

4. Kitcher's ideal of deliberation

The question, then, is whether Kitcher's own ideal is more realistic than the value-free ideal. So how does his new ideal look like? To start with, Kitcher's division of epistemic labor differs from the traditional model in that citizens participate not only in scientific agenda-setting ("investigation") and the application of findings ("transmission"), but also in the determination of scientific standards ("submission") and the acceptance of results ("certification"). Crucially, however, citizens are to do so in a specific way: one that aligns with the ideal of deliberation.

To see how this is supposed to work, remember that Kitcher views ethics as a "social technology", i.e. as an expression of humanity's evolutionary attempt to foster cooperation and social cohesion (ibid., pp. 41-49). To achieve this, Kitcher holds, humanity has developed an effective mechanism: group deliberation. Whenever important issues had to be settled, "our Paleolithic predecessors sat down together to decide on the precepts for governing their group life" (ibid., p. 43). In these discussions "all adult members of the band are to be heard, and the wishes of each must be considered" (ibid., p. 43). This helped to coordinate activities, which in turn ensured the "survival of the group" (ibid.). However, deliberation is not merely a remnant of our evolutionary past. Until today, Kitcher holds, deliberation is "[t]he only vehicle available to us" (ibid., p. 50) to define ethical standards that are acceptable to all. Kitcher's aim is thus to transfer the deliberative approach to current ethical problems, including the determination of value-laden choices in science.

Now, the concept of deliberation involves certain procedural requirements. It is these requirements that make deliberation an *ideal*. Five such requirements can be identified in Kitcher's book:

* Inclusion: *ideal deliberation includes everyone who is, or may potentially be, affected by the topic under consideration*. This requirement employs what is known as the all-affectedness principle (Dewey, 1927; Brown, 2009) and rests on Kitcher's conviction that the deliberative mechanism invented by our Paleolithic ancestors should now be extended beyond the local community. As in a globalized world everyone is potentially affected by a decision, Kitcher's ideal moves "from the small band to the human population" (ibid., p. 54), such that "the body of discussants [comprises all] members of our species, including those who will come after us" (ibid., p. 50).
* Equality: *ideal deliberators enjoy equal deliberative rights and opportunities.* This requirement follows from Kitcher's claim that ethical standards are "not for any single person […] to determine" (ibid., p. 111). Rather, the only way "for arriving at judgments about values is discussion in which the participants come as equals" (ibid.). All deliberators are thus to receive equal authority, and this equality must not be undermined by hierarchy, social status, or power.
* Impartiality: *ideal deliberators view their own desires on par with those of anyone else.* This requirement demands that, for one, "the perceived desires of those with whom one deliberates are given equal weight with one's own" (ibid., p. 51) and that, for another, deliberators acknowledge the perspective that any given deliberator has on the desires of any other deliberator. This "mirroring", as Kitcher calls it, makes it that egocentric desires are "filtered" (ibid.) out, such that only the universally acceptable desires survive.
* Understanding: *ideal deliberators understand the discussed topic to the degree necessary to anticipate the consequences of a given proposal on themselves and others*. This requirement is decisive, as "even the most well-disposed conversationalists will arrive at peculiar recommendations if they are in the grip of flawed ideas" (ibid., p. 51). Kitcher's ideal hence presupposes that deliberators do "not rely on false beliefs about the natural world" (ibid.).
* Rationality: *ideal deliberators judge proposals on grounds of defensible reasons and are ready to change their beliefs accordingly*. While Kitcher is less explicit about this requirement, it follows from his "mirroring" and from the fact that he assumes a phase of "tutoring" (ibid., p. 128) where deliberators learn about the current state of science. Both mirroring and tutoring presuppose a type of rationality, namely an orientation towards giving and taking reasons.

Now, it seems obvious that full compliance with these requirements is not particularly realistic. This is most evident for the inclusion requirement. There is no way to deliberate with the entire human species, let alone with "those who will come after us" (ibid., p. 50). What is more, science consists of countless individual research projects, and if value-judgements really occur throughout any given project, as Kitcher claims, no deliberative body can discuss all of these judgements. Similar things can be said about the other requirements. For instance, scholars have long argued that deliberations often fail to realize equality, e.g. because they presuppose discussion skills that underprivileged groups typically lack (Sanders, 1997; Young, 2001). Deliberations are also susceptible to group-think (Sunstein, 2006), which makes it unlikely that deliberators meet the impartiality requirement, particularly when it comes to the interests of outgroups. It also seems unlikely that deliberators meet the understanding requirement, given the lack of scientific literacy in the public (NSB, 2016) and the proneness of every-day reasoning to formal errors (Kahneman et al., 1982). Finally, the fact that people tend to ignore, or even reject, unwelcomed evidence (Kahan et al., 2013; Lewandowsky & Oberauer, 2016) makes the rationality requirement seem highly unrealistic.

This seems to be a major problem for Kitcher: on the one hand, he rejects the value-free ideal for being unrealistic; on the other hand, his alternative does not seem particularly realistic itself. One way to address this problem would be to maintain that ideal deliberation is less unrealistic than it appears. But this is not what Kitcher does. On the contrary, he acknowledges that "[a]ctual deliberations are often, probably always, infected by special interests, ideological presuppositions, and inequities" (2011, p. 125). If this is so, then Kitcher is either inconsistent – or his ideal invulnerable to the realist challenge *despite* its lack of realism. I will argue in the following that the answer to this question depends on what notion of "ideal" we are referring to.

5. Four notions of "ideal"

To see how we should conceive of Kitcher's ideal, I propose to distinguish four types of ideals:

* A *blueprint ideal* is a representation of a normatively eminent state of affairs that we both should and can realize, although the efforts necessary to do so may be considerable. An example may be "students of philosophy should read all three of Kant's critiques".
* A *compass ideal* is an indicator that steers our activities into a normatively desirable direction. Contrary to a blueprint, a compass does not compel us to realize a specific state of affairs, but to proceed as far as possible on the indicated path. As an illustration, take "students of philosophy should read the entire primary and secondary literature on Kant".
* A *yardstick ideal* is a standard that evaluates an actual state of affairs by measuring it against an ideal state, such that any state that has similar normative features as the ideal is deemed similarly valuable as the ideal. Contrary to a blueprint and a compass, yardsticks are flexible regarding the way in which the ideal is realized. An example may be "students of philosophy should do whatever promotes their philosophical skills equally well as reading Kant's critiques".
* A *reconstructive ideal* is a representation of a desirable state of affairs, the realization of which is implicitly presupposed in a given practice. Contrary to the other notions, reconstructions do not only regulate a practice, but constitute it in the first place. As an example, take "students of philosophy should follow the implicit norms that are constitutive for philosophy as a social practice (e.g. knowledge of the philosophical literature, careful argumentation etc.)".

Note that I neither claim that this distinction covers all possible notions of "ideal", nor that I am the first to explore the conceptual issues that underlie these notions (see e.g. the so-called ideal/non-ideal theory debate in political philosophy, Simmons, 2010; Hamlin & Stemplowska, 2012; Valentini, 2012). However, the distinction helps to understand Kitcher's deliberative ideal and may, in addition, offer an analytical lens for the axiological turn more generally (e.g. by grasping the difference between different types of critiques of the value-free ideal). Another clarification is that the four interpretations are not mutually exclusive. For instance, a student who attempted to read Kant's critiques (a blueprint ideal), but finds that this was overambitious, may still read as much

of these books as possible (a compass ideal). Similarly, the skills that a student acquires by reading the three critiques may provide a standard to judge the skills of someone who studied, say, formal logic (a yardstick ideal). For the Kant example to work in the reconstructive notion, however, one would have to argue that knowledge of Kant's critiques is part of what constitutes philosophy as a social practice. While this does not seem particularly convincing, it still illustrates how reconstructive ideals work in general: they point towards implicit norms of a social practice, arguing that the practice would cease to exist, or turn to something else, if these norms were consistently violated.

As a final remark, note that the above typology takes up, but also goes beyond the common distinction between *regulative* and *constitutive* norms (Rawls, 1955; Searle, 1995). Regulative norms are prescriptions for activities that are conceptually independent from these rules, while constitutive norms are preconditions of the activities themselves. John Searle has famously exemplified the former with traffic regulations and the latter with chess rules: driving on the wrong side of the road makes you a bad driver, but it does not undermine your status as a driver. But if you use your rook like a knight, you are not merely a bad chess player – you rather cease to play chess (1995, p. 27-29). Now, in my typology, blueprint, compass and yardstick ideals are regulative, while reconstructive ideals are constitutive. Yet, the typology goes beyond the classic distinction, as it specifies different types of regulative ideals. Additionally, the reconstructive notion differs from Searle's chess example in that the question whether the ideal has been attained cannot be answered by some external authority (say, a referee), but must be answered within the practice itself. Furthermore, I will show later that reconstructive ideals are more implicit than, e.g., chess rules, and may even be *embryonic* in the sense that they constitute a practice without being fully realized. In this respect, the reconstructive notion stands more in the tradition of Hegel and critical theory than in Searle's.

6. Deliberation: a blueprint, a compass, or a yardstick?

6.1 THE BLUEPRINT NOTION: IDEAL DELIBERATION AND THE ARGUMENT FROM NON-BINDINGNESS

Returning to Kitcher's ideal, let us now see how the realist challenge affects a blueprint, compass, or yardstick interpretation of ideal deliberation. For the sake of simplicity, I focus on one realist argument for each notion (this does not mean that these notions are not susceptible to the other realist arguments, but simply that some argument-notion pairs are more straightforward than others). Starting with the blueprint notion, the argument from non-bindingness offers the most promising option for a realist critic. This is because blueprint ideals are supposed to be attainable, which is exactly what Kitcher's ideal does not seem to be. In fact, much seems to speak against even considering this notion in our context, be it Kitcher's own claim that his ideal will *never* be fully realized (2011, p. 54, p. 115, p. 125), or be it other voices in the literature who view Kitcher's ideal more as a yardstick (Douglas, 2013; Keren, 2013) or a compass (Wilholt, 2014), but clearly not as a blueprint.

Yet, there are still reasons to consider the blueprint notion in our context. First, it has merits that the other notions lack, such as its specificity (e.g. "read Kant's three critiques") or the plausible claim that some responsibilities cannot be fulfilled by mere approximation (e.g. some policy goals of the United Nations are highly ambitious, yet we are supposed to fully achieve them). Second, Kitcher himself uses the blueprint notion against the value-free ideal, arguing that scientists cannot be held accountable for not reaching an unreachable goal (2011, p. 31). It is thus fair to ask whether his own ideal would stand the same test. And third, it turns out that interpreting Kitcher's ideal as a blueprint is not so implausible after all, at least with respect to some of its requirements.

Consider Kitcher's equality and understanding requirements. I said before that, given the *de facto* inequalities in current societies and the widespread scientific illiteracy, these requirements do not seem particularly realistic. In fact, no-one doubts that real-world deliberations often fall short of equality and understanding. But are these problems really *insurmountable*? Equality can be promoted by various measures, including the admittance of less "privileged" forms of expression (Bächtiger et al., 2010), regulated

speaking times, sanctions for intimidating behavior, or the involvement of professional moderators (see e.g. ECAST, 2015). Understanding can be promoted by what Kitcher calls "tutoring" (2011, pp. 111-125), i.e. a phase where deliberators learn about scientific methods and results, or by experts "on stand-by" who clarify technical aspects during the discussion (e.g. ECAST, 2015). It has also been shown that participants are surprisingly well-informed after a deliberation (e.g. Esterling et al., 2011; Fournier et al., 2011, p. 37). Note, however, that I do not claim that understanding and equality will always be achieved. My point is rather that these requirements do not seem to be *intrinsically* unattainable. This is a crucial difference to Kitcher's arguments against value-freedom, where the claim is that *no effort whatsoever* will suffice to reach this ideal. I do not think that this is the case for equality and understanding. Hence, the argument from non-bindingness is not successful against these requirements (they may, of course, still be subject to the remaining realist arguments, but I focus on the bindingness issue here).

Regarding impartiality and rationality, I just said that blueprints have the merit of specificity, i.e. they allow for concrete goal formulations. Yet it is doubtful whether this advantage can be cashed out when it comes to these two requirements. This is because it can, and typically will, be opaque to us whether a deliberator really weighs all interests equally (impartiality) and is motivated by defensible reasons (rationality). A situation where these requirements are violated can thus be indistinguishable from one where they are fulfilled. While this is not a specific problem of the blueprint notion, it is more unsatisfying here, as it undermines one of the main reasons that make blueprints appealing. On the other hand, however, I would still argue that this does not force us to reject a blueprint interpretation of Kitcher's ideal, at least not by way of the argument from non-bindingness. After all, attainability and applicability are two different things – just because we do not know whether a goal has actually been achieved does not mean that the goal is essentially unachievable. And as long as there is no in principle argument similar to Kitcher's arguments against value-freedom, it seems unwarranted to assume that impartiality and rationality are unattainable.

What I said so far implies that, if we only consider equality, understanding, impartiality and rationality, Kitcher's claim that reality "probably always" (2011, p. 125) falls short of his ideal may have been too concessive. However, things are different when it comes to the inclusion requirement. I said before that it is impossible to have all humanity, let alone future generations, be part of the "body of discussants" (ibid., p. 50). The only way to engage with the whole "range of perspectives found in the inclusive human population" (ibid., pp. 53-54) is through *representation*. Also, as it is impossible to hold delib-

erations about each and every aspect of each and every research project, the discussed issues have to be *generalized*, such that deliberators discuss types of research decisions rather than individual choices. And indeed, representation and generalization make the ideal realistic, thus avoiding the argument from non-bindingness. The trouble, however, is that they do so in ways proposed by the compass or yardstick notion, but not the blueprint notion.

Regarding generalization, this is because whatever result a deliberation may have, it needs to be *interpreted* in order to be applicable to a given research choice. To apply a general guideline to a specific case is to hold that the deliberation *would* endorse this interpretation if it *were* held in this case. Such counter-factual reasoning is compatible with the yardstick notion, but not with the blueprint notion. Representation, on the other hand, is compatible with the compass notion if the aim is to consider as many social perspectives as possible, and compatible with the yardstick notion if the aim is to determine what those not present *would* have contributed if they *had* participated. Universal inclusion is therefore either impossible, in which case the requirement is normatively non-binding, or it must be interpreted in a non-blueprint manner.

6.2 THE COMPASS NOTION: IDEAL DELIBERATION AND THE ARGUMENT FROM UNINTENDED EFFECTS

So if the blueprint notion of Kitcher's ideal falls prey to the argument from non-bindingness, why not embrace a notion that is *qua* design invulnerable to this argument? The compass notion seems to be the obvious choice here, as it only requires us to *approximate* a certain state of affairs ("the closer, the better"). Since approximations are typically attainable, this addresses the bindingness problem and may even increase the ideal's practical relevance by avoiding the impression of a lost cause. It is thus unsurprising that Kitcher draws heavily on this notion (2011, pp. 45-49, p. 54, pp. 125-130, pp. 223, see also Wilholt, 2014). Kitcher argues that "[a]lthough we cannot hope to live up to" the ideal, it still indicates "*directions* in which *actual* conversations about values might proceed" (2011, p. 54, orig. italics). Kitcher also emphasizes that approximations of his ideal already exist in the form of deliberative citizen panels (ibid., p. 222; Fishkin, 2009). While only partial realizations, Kitcher argues, this shows that something close to ideal deliberation is in fact feasible.

As it turns out, however, the compass notion of Kitcher's ideal is vulnerable to the argument from unintended effects, or rather a version of this argument that is known as the

problem of the second best (Lipsey & Lancaster, 1956). It maintains that if a goal is found to be unachievable, the second best solution may not be to approximate, but to give up on that goal. A nice illustration is Kim Stanley Robinson's novel "Aurora", where a spaceship returns to Earth without the necessary fuel to decelerate. The ship's AI thus calculates a trajectory through the solar system, using the gravity of the Sun and the planets to slow down. The trouble is that the slightest error will result in the ship either crashing into one of these bodies, or being catapulted out of the system. In this situation, the next best thing to a perfect trajectory is not some approximation, but not to embark on that journey in the first place (the crew actually had this alternative). As Robinson puts it: "in certain problems, only 100 percent will do; 99 percent is still a complete miss" (2015, p. 390).

So is Kitcher's ideal is prone to similar second-best problems? Here, I would again argue that the issue is not so much equality, impartiality, understanding, or rationality, but once more the inclusion requirement. With respect to the former four requirements, it seems plausible that if there are two states of affairs *s* and *s**, and if *s** exhibits a greater proximity to the ideal than *s*, then *s** is preferable to *s*. Someone who embraces Kitcher's ideal will therefore generally prefer a more equal, impartial, insightful, and rational deliberation over one that is less equal, impartial, insightful, and rational. Hence, contrary to the predicament of Robinson's star travelers, the next best thing to a full realization of these requirements is to approximate them. As it turns out, however, this "the closer, the better" logic is problematic when it comes to the inclusion requirement.

To see this, remember that the ideal "takes the body of discussants to be the members of our species" (2011, p. 50). If we take this literally, the compass notion implies that larger deliberative bodies are generally preferable to smaller ones. But this is clearly false, not the least because there is a systematic tension between a deliberation's size and its capacity to realize the other four requirements: the larger the deliberation, the less likely is it that everyone gets heard (equality), that everyone mirrors the desires of everyone else (impartiality), that everyone receives proper tutoring (understanding), and that everyone considers all available reasons (rationality). This is supported by other, more external considerations, e.g. that large deliberations are time-consuming and expensive, thus exhausting resources that could be used for other valuable projects. Thus, contrary to the "the closer, the better" logic, more inclusion will *at some point* be worse than less inclusion.

An obvious rejoinder to this reasoning is that the requirement can be understood in a less literal sense, namely as inclusion of *social perspectives* rather than of *individuals*. As members of social groups tend to share similar experiences, their perspectives can be represented by a relative small number of representatives (Brown, 2006; Fishkin, 2009; ECAST, 2015). One way of selecting these representatives is to determine demographic criteria and then make random draws from the resulting demographic pools[5]. For instance, a deliberation organized by the Danish Board of Technology gathered about 10.000 citizens, split into 97 meetings in 76 countries (DBT, 2015). The participants were chosen to "reflect the demographic distribution in their country or region with regards to age, gender, occupation, education and geographical zone of residency" (ibid., p. 15). In addition to such demographics-based representation, Kitcher suggests to represent those "who cannot speak for themselves" (2011, p. 132) – children, people with serious diseases, future generations –by including "people who know them intimately and who are devoted to their interests" (ibid.).

But this does not solve the problem that more inclusion is not always better. This is because there is a potentially endless number of social perspectives, reflected by innumerable demographic criteria and combinations of criteria. Even the huge deliberation of the Danish Board of Technology hardly represented the entire "range of social perspectives found in the inclusive human population" (ibid., pp. 53-54), let alone every perspective that future generations may have. In order to include representatives of all these perspectives, a deliberation would still have to be so large that more inclusion will *at some point* undermine the other four requirements and generate unreasonably high costs. The problem of the compass notion thus remains: there is always a state of affairs that approximates universal inclusion more closely, but this state is not generally preferable to a less inclusive state of affairs. A realist critic can therefore use the argument from unintended effects to argue that ever closer approximations of Kitcher's ideal would lead to undesired consequences.

5 As Heather Douglas (2013) has rightly noted, demographic representation is just one of many ways in which social perspectives may be represented in a deliberation (see also Brown 2006, 2009). I will not get into this discussion because, first, Kitcher himself seems to have something like demographic representation in mind (2011, pp. 222-224) and, second, demographic representation is the most common approach in actual citizen deliberations (see e.g. Fishking, 2009; DBT, 2015; ECAST, 2015).

6.3 THE YARDSTICK NOTION: IDEAL DELIBERATION AND THE ARGUMENT FROM IRRELEVANCE

Interestingly, Kitcher himself acknowledges these problems. He argues that "any attempt to orchestrate even a sample of voices representative of the diverse perspectives of living people would produce a vast cacophony" (2011, p. 51), which in turn "would doom any chance of serious discussion" (ibid.). Yet, Kitcher has a solution at hand. He proposes to conduct deliberations in a *counter-factual* manner, such that "the ethical conclusions to be endorsed are those that *would* emerge from an ideal conversation" (ibid., p. 51, my emphasis, see also ibid., p. 106). Kitcher illustrates this approach by outlining how an ideal deliberation would presumably react to several issues in current science (ibid., pp. 230-248). This has led some authors to call Kitcher's ideal a "thought experiment" (Douglas 2013, p. 903) or a "hypothetical procedure" (Keren 2013, p. 238). As it turns out, however, this view of ideal deliberation runs into problems with the argument from irrelevance.

The hypothetical approach rests on what I call the yardstick notion. It differs from the notions discussed so far, as it focusses on the *content* of a decision or a judgement rather than the *procedure* by which it is determined. The function of ideal deliberation is hence to measure, not to produce decisions or judgments. As Kitcher puts it: "[o]ur ethical discussions are adequate to the extent they reach the conclusions that would have resulted from an ideal deliberation" (ibid., p. 51). More generally speaking, yardstick ideals capitalize on the fact that two states of affairs can be phenomenologically different and yet exhibit similar normative features. Reading Kant and studying formal logic, for instance, are two different things, but they may be equally desirable with respect to their philosophical value. Similarly, conducting a thought experiment is not the same as holding a deliberation, but perhaps it is *just as good* with respect to the value-judgements it creates.

However, the applicability of a yardstick depends on whether it provides a sufficiently clear benchmark. If we did not know what skills a student acquires by reading Kant, it would be opaque to us whether studying logic will yield equally valuable results; and if we did not know what conclusion a deliberation would endorse, it would be unclear whether conducting a thought experiment is just as good as holding an actual deliberation. Kitcher himself is somewhat ambiguous on this point. In certain cases, he argues, the outcome of an ideal deliberation "is easily predicted" (ibid., p. 234, see also pp. 135-136), whereas in other cases "nobody can predict how the ideal conversation would come to conclusion" (ibid., p. 124, see also p. 176). If it is true, however, that yardsticks

need clear benchmarks, the latter cases are problematic. Just as critics of value-freedom have claimed that "science, even science done well, is not value-free, and so the ideal is irrelevant" (Douglas, 2015, p. 5), a critic of Kitcher's ideal can claim that a thought experiment, even if done well, cannot replace actual deliberation – and so the yardstick notion of Kitcher's ideal is irrelevant.

One possible rejoinder would be that those cases where deliberation outcomes are predictable are common enough to warrant the yardstick notion. Unfortunately, this rejoinder fails. First, it is unlikely that anyone can anticipate all reasons and perspectives that would be available in a large-scale deliberation. As Heather Douglas has noted, Kitcher's ideal may not "survive the lack of actual conversational instantiation" (2013, p. 903), as "it is extremely difficult to see all the important factors [that] should be part of the conversation" (ibid., p. 905). Second, we just cannot foresee the result of a process where all deliberators take the perspective of all other deliberators, as the impartiality requirement demands. It is, e.g., not so clear that mirroring religious and non-religious perspectives would, as Kitcher claims (2013, p. 234), filter out the religious perspectives. Finally, even if the deliberators' verdict on some issues were predictable, these issues would still be part of a chain of less predictable issues. Prioritizing pressing social problems over scientific curiosity, for instance, may be an obvious choice for deliberators (ibid., p. 123), but this leaves unspecified what makes a problem "pressing", how social problems should be prioritized among each other, or what research approach should be used to tackle them. It is these kinds of questions that underlie actual science, and it seems impossible to know in advance how a deliberation would answer them.

The problem with using Kitcher's ideal as a yardstick is thus the lack of a clear benchmark. Inclusion and impartiality are most problematic in this regard, as these requirements imply an unpredictable interplay of desires and reasons, mirrored by a potentially infinite number of social perspectives. This triggers the argument from irrelevance: how can an ideal be an effective means of coordination – a "social technology", as Kitcher has it – if we do not even know what the ideal requires us to do? Rather than being an actual decision factor, it seems more likely that scientists would project their own value-judgements into the ideal. Similar to some critiques of value-freedom (e.g. Longino, 2002), one can then argue that it may appear on the surface as if scientific choices were driven by Kitcher's ideal, whereas in reality other factors were dominant. Thus, if Kitcher wishes to maintain the standard direction of fit, he has to consider an alternative notion of ideal deliberation.

7. Deliberation as a reconstruction

7.1 CONCEPTUAL BACKGROUND

I have argued that Kitcher's ideal falls prey to the realist challenge if interpreted in a blueprint, compass or yardstick manner. One way to address this problem may be to modify the ideal, e.g. by adopting a less demanding version of the inclusion requirement. Yet, this would amount to what I have called a reversion of the direction of fit, i.e. the ideal would be fitted to the world rather than the other way around. Before making such a concession, a supporter of Kitcher's ideal will want to make sure that there is really no way to keep the more ambitious version of the ideal. As it turns out, however, there is indeed such a way: the reconstructive notion.

Reconstructive ideals differ from the other notions in that they are *constitutive*, i.e. they "do not merely regulate, they also create the very possibility of certain activities" (Searle, 1995, p. 27). Whereas a failure to comply with a blueprint, a compass or a yardstick is just blameworthy, failing to comply with a reconstructive ideal sheds doubt on whether we are actually performing the activity in question. Apart from Searle's regulative/constitutive distinction (see also Rawls, 1955), the reconstructive notion reaches back to Hegel, who held that the philosopher's task is to articulate the implicit ethical order or *Sittlichkeit* of a given historical context (1807/2017). The idea that ethical orders reside, perhaps in an embryonic stage, in the fabric of our social world has later influenced Marx, Habermas, and recent critical theory (Jaeggi, 2018). The reconstructive notion takes up this tradition, as it does not prescribe norms from the outside, but expresses the implicit standards that constitute a practice. By "locat[ing] the normativity of social practices in the performance conditions of these practices themselves" (Jaeggi, 2018, p. 190), the notion makes it possible to "criticiz[e] the object in terms of a standard that lies in the object itself" (ibid., p. 181)[6].

6 Rahel Jaeggi distinguishes two types of critique that match up with what I call the yardstick notion: a less ambitious *internal* critique that "seeks to reinstate the principles that make up the life of a community or to reactivate the real meaning of its ideals" (2018, p. 181), and a more ambitious *immanent* critique that "starts from […] problems and moments of crisis internal to a form of life" (ibid., p. 31) and then "transcends this starting point" (ibid.) by initiating "a transformation of forms of life" (ibid.). While I will not discuss this distinction here, I would argue that yardstick ideals can be used in both of these ways, depending on the specific ideal under consideration (e.g. my philosophy example seems to be more internal in Jaeggi's sense, whereas the deliberative ideal seems to amount to a more immanent form of critique).

To better understand the reconstructive notion, note that the implicit norms of a practice need not be easily interpretable. For instance, the practice of philosophy involves certain standards of reasoning whose exact meaning is debatable; but this does not mean that they do not exist, or that they are irrelevant in practice. Another aspect is that, as such standards reside in the practice itself, the question of whether they are violated must be decided within the practice and cannot be determined by some external authority (the reconstructive notion differs from Searle's chess example here). Furthermore, reconstructive ideals can be *embryonic* in the sense that they are implicit in a practice without being fully realized. Philosophers may, e.g., sometimes fail to be, say, critical or careful, but this neither refutes these norms, nor the fact that critical and careful reasoning is part of what it means to be a philosopher (this is another difference to Searle's chess example).

Transferring this to Kitcher's ideal, one may wonder what exactly is being *reconstructed* in a reconstructive notion of ideal deliberation. A promising answer to this question can be found in the early works of Jürgen Habermas (1979, 1984) and Karl-Otto Apel (1980). Habermas famously proposed a concept of *communicative rationality*, arguing that our every-day linguistic practice essentially consists in exchanging certain validity claims. For instance, by saying "please, shut the door" a speaker implicitly claims that shutting the door is in fact possible, that her demand is normatively acceptable, that she expresses an authentic wish, and that her speech act is actually comprehensible (see Habermas, 1984, p. 306). However, in a communicative or non-strategic context a hearer can always *reject* these validity claims, e.g. by arguing that the door is stuck. The speaker, on the other hand, can reject this rejection, say, by arguing that the door is not stuck at all. This can spark a process of *giving and taking reasons*, where with every new speech act "the speaker proffers a speech-act immanent obligation to provide justification" (Habermas, 1979, p. 64). Of course, such an exchange of reasons will not always unfold, but this does not change the fact that, at least in a communicative context, we must be ready to provide reasons should they be demanded.

Apel (1980) has called these implicit obligations the "*a priori* of the unlimited communication community" (ibid., p. 267). He held that the justifications we give, or could give, for our linguistic and other behavior are not only directed towards our actual interlocutors, but towards "an *ideal communication community* that would basically be capable of adequately understanding the meaning of [our] arguments and judging their truth" (ibid., p. 280, orig. italics). This community of ideal deliberators is unrealistic in the sense that it does not actually exist; yet, it is realistic in a different sense, namely that it is implicitly present in our linguistic practices. As Apel puts it, "the ideal community is

presupposed […] *in* the real one, namely, as a real possibility of the real society, although the person who engages in argument is aware that (in most cases) the real community, including himself, is far removed from being similar to the ideal community" (ibid., pp. 280-281, orig. italics).

7.2 THE RECONSTRUCTIVE NOTION AND THE REALIST CHALLENGE

Now, my proposition is to interpret the deliberative ideal in exactly the same way: as a set of implicit norms that we presuppose whenever we exchange validity claims in a communicative setting. Kitcher's requirements can then be seen as the articulation of a speaker's implicit claim that she could justify a speech act not only to a group of actual hearers, but to an ideal deliberative community that comprises all types of social perspectives (inclusion), that judges proposals in a non-hierarchical way (equality), that gives everyone's desires equal weight (impartiality), that understands the topic at hand (understanding) and that is motivated by intersubjective reasons (rationality). Vice versa, the speaker implicitly claims that she will assess any replies to her speech act in a way that complies with these requirements. Whether these implicit claims can be maintained, however, cannot be decided from the outside, but must be judged in the communication setting itself. A hearer can always reject a speech act, arguing that one or more of the implicit claims fails. Moreover, when a speaker *constantly* violates communicative norms, this will shed doubt on whether she actually does what she claims to do – engaging in the practice of giving and taking reasons.

But can the reconstructive notion not be refuted by the same arguments that troubled the other notions of Kitcher's ideal? With regard to the argument from non-bindingness, it is true that any actual exchange of validity claims will be "far removed from being similar to the ideal community" (Apel, 1980, p. 281). Yet, in contrast to the blueprint notion, the reconstructive notion does not prescribe an *ought*, but only articulates it. If this is so, then the ethical order that is implicit in our communicative practices cannot be given up as an ideal, as it manifests itself whenever we give and take reasons. Vice versa, cases of serious violations of these norms do not refute the ideal, but merely illustrate that there are linguistic practices besides genuine communication (bargaining, exercising social power etc.). This reveals an interesting twist: although the reconstructive notion of Kitcher's ideal is essentially unattainable, it cannot be refuted via ought-implies-can, simply because there cannot be an *ought* to give up on the ideal if we *cannot* do so.

Regarding the argument from unintended effects, I concede that the reconstructive notion of ideal deliberation has a tendency towards ever more inclusion, just as the compass notion has. But it is not so clear that this implies a general preference for larger deliberative bodies. After all, we constantly exchange reasons in our every-day linguistic practices, and the implicit claim that these reasons would stand the scrutiny of an ideal community does not mean that this community, and all the social perspectives it comprises, must be present here and now. It rather seems to suffice that we defend our claims against arguments from a different perspective *once these arguments are brought up*. For this to be possible, however, there must be opportunities for representatives of different social perspectives to enter the conversation. Consequently, there is always a *caveat* to the social composition of a deliberative body. Such a body must remain open towards outsiders, such that perspectives that have not been considered in the original choice of participants can still enter later in the process. If this is given, the inclusion requirement seems to be *prima facie* fulfilled.

With respect to the argument from irrelevance, the problems of the yardstick notion are much less pressing in the case of the reconstructive notion. While the reconstructive view of ideal deliberation does involve a counter-factual element – the assumption that a validity claim *could* be justified to a community of ideal deliberators if they *would* scrutinize the claim – it does not involve the idea we need not engage in actual conversations. On the contrary, the reconstructive notion is intimately linked to actual communicative practices. A reconstructive ideal does not, as a yardstick would, start from a normative principle and then apply this principle to a practice; rather, it starts from a practice and asks whether the practice complies with its own constitutive norms. In order to be an effective means of social coordination, the reconstructive approach to deliberation is therefore be to engage in actual conversations, e.g. in the form of mid-scale citizen panels, and exchange the reasons that are relevant to judge a given topic. It will then be unnecessary to know in advance what the result of such a conversation would be, and whether it complies with the ideal's requirements. Rather, we this judgement would be left to the deliberators themselves.

8. Conclusion

Philip Kitcher's ideal of deliberation is among the most relevant contributions to the ongoing debate on values in science. However, this debate has not always been clear regarding the specific meaning of the term "ideal". In this paper, I have proposed four interpretations of this term: the *blueprint*, the *compass*, the *yardstick*, and the *reconstructive* notion. I have emphasized the conceptual differences between these notions, showing that each of them offers different options for defending and refuting a given ideal. This is particularly relevant when it comes to unrealistic ideals such as the one proposed by Kitcher. I have argued that unrealistic ideals are subject to a *realist challenge*, i.e. the claim that an unattainable goal is (i) normatively non-binding, (ii) prone to unintended effects, and (iii) irrelevant in practice. Whereas a valid ideal would instantiate what I have called the *standard direction of fit* (where the world is supposed to fit the ideal), an ideal that is refuted by the realist challenge would have to *reverse the direction of fit* (such that the ideal is supposed to become more similar to the world). I have argued that the most promising way to keep the standard direction of fit in the case of Kitcher's ideal is the reconstructive notion. By locating the normativity of Kitcher's ideal in the actual practice of exchanging validity claims, the ideal is realistic and ambitious at the same time. In addition to offering an analytical lens for Kitcher's approach, the four notions contribute to the axiological turn more generally. Differentiating between different types of ideals may be helpful for discussing several issues in this debate – not only alternatives to the value-free ideal, but also the traditional value-free ideal itself. Whether the ideal of value-freedom can be defended more effectively when interpreted in the reconstructive notion remains to be seen.

Acknowledgements

This work was funded by the German Research Foundation (DFG) (254954344/GRK2073). Additional funding was provided by the German Academic Exchange Service (DAAD). I would like to thank Philip Kitcher, Torsten Wilholt, Dietmar Hübner, Jan-Philipp Kruse, Hauke Behrendt, Mirko Suhari and the members of the research training group GRK 2073 "Integrating Ethics and Epistemology of Scientific Research" for valuable feedback. This paper also benefited from discussions with members of the Consortium for Science, Policy and Outcomes at Arizona State University and members of the Mercator Research Institute on Global Commons and Climate Change, particularly Martin Kowarsch.

References

Anscombe, G. E. M. (1957). *Intention*. Blackwell.

Apel, K.-O. (1980). *Towards a transformation of philosophy*. Routledge.

Bächtiger, A., Niemeyer, S., Neblo, M., Steenbergen, M. R., & Steiner, J. (2010). Disentangling Diversity in Deliberative Democracy: Competing Theories, Their Blind Spots and Complementarities. *Journal of Political Philosophy*, 18(1), 32–63. https://doi.org/10.1111/j.1467-9760.2009.00342.x

Biddle, J. (2013). State of the field: Transient underdetermination and values in science. *Studies in History and Philosophy of Science*, 44(1), 124–133. https://doi.org/10.1016/j.shpsa.2012.09.003

Bright, L. (2018). Du Bois' democratic defence of the value free ideal. *Synthese*, 95(5), 2227–2245. https://doi.org/10.1007/s11229-017-1333-z

Brown, M. B. (2006). Survey Article: Citizen Panels and the Concept of Representation. *Journal of Political Philosophy*, 14(2), 203–225. https://doi.org/10.1111/j.1467-9760.2006.00245.x

Brown, M. B. (2009). *Science in democracy. Expertise, institutions, and representation*. MIT Press.

Brown, M. J. (2013). Values in Science beyond Underdetermination and Inductive Risk. *Philosophy of Science*, 80, 829–839. https://doi.org/10.1086/673720

Bueter, A. (2015). The irreducibility of value-freedom to theory assessment. *Studies in History and Philosophy of Science*, 49, 18–26. https://doi.org/10.1016/j.shpsa.2014.10.006

Danish Board of Technology Foundation (DBT) (2015): *World Wide Views on Climate and Energy. From the world's citizens to the climate and energy policymakers and stakeholders*. http://climate-andenergy.wwviews.org/wp-content/uploads/2015/09/WWviews-Result-Report_english_low.pdf

Dewey, J. (1927): *The Public and Its Problems*. Swallow Press.

Douglas, H. E. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, 67(4), 559–579. https://doi.org/10.1086/392855

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Douglas, H. E. (2013). Review: Philip Kitcher. Science in a Democratic Society. *British Journal for the Philosophy of Science*, 64(4), 901–905. https://doi.org/10.1093/bjps/axt006

Douglas, H. E. (2015), Values in Science. In P. Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science* (pp. 609–630). Oxford University Press.

Elliott, K. C. (2011). *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. Oxford University Press.

Esterling, K. M., Neblo, M. A., & Lazer D. M. J. (2011). Means, Motive, and Opportunity in Becoming Informed about Politics: A Deliberative Field Experiment with Members of Congress and Their Constituents. *Public Opinion Quarterly*, 75(3), pp. 483–503. https://doi.org/10.1093/poq/nfr001

Estlund, D. (2014). Utopophobia. *Philosophy & Public Affairs*, 42(2), 113–134. https://doi.org/ 10.1111/papa.12031

Expert and Citizen Assessment of Science and Technology Network (ECAST) (2015): *Informing NASA's Asteroid Initiative: A Citizen's Forum*. https://ecastnetwork.org/research/informing-nasas-asteroid-initiative-a-citizens-forum

Fishkin, J. S. (2009). *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press.

Fournier, P., van der Kolk, H., Carty, R. K., Blais, A., & Rose, J. (2011). *When citizens decide: Lessons from citizen assemblies on electoral reform*. Oxford University Press.

Geden, O., & Beck, S. (2014). Renegotiating the global climate stabilization target. *Nature Climate Change*, 4, 747–748. https://doi.org/10.1038/nclimate2309

Geden, O. (2015, December 14). *Paris climate deal: the trouble with targetism*. The Guardian. https://www.theguardian.com/science/political-science/2015/dec/14/the-trouble-with-targetism

Gregory, A. (2012). Changing Direction on Direction of Fit. *Ethical Theory and Moral Practice*, 15(5), 603–614. https://doi.org/10.1007/s10677-012-9355-6

Guillemot, H. (2017). The necessary and inaccessible 1.5°C objective. A turning point in the relations between climate science and politics? In S. Aykut, J. Foyer, & E. Morena (Eds.), *Globalising the Climate. COP21 and the Climatisation of Global Debates* (pp. 39–56). Routledge.

Habermas, J. (1970): The Scientization of Politics and Public Opinion. In *Toward a Rational Society* (pp. 62–80). Beacon Press.

Habermas, J. (1979). What is Universal Pragmatics? In *Communication and the Evolution of Society* (pp. 1–68). Beacon Press.

Habermas, J. (1984). *The Theory of Communicative Action. Vol. 1: Reason and the Rationalization of Society*. Beacon Press.

Hamlin, A., & Stemplowska, Z. (2012). Theory, Ideal Theory and the Theory of Ideals. *Political Studies Review*, 10(1), 48–62. https://doi.org/10.1111/j.1478-9302.2011.00244.x

Harding, S. (1995). "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3), 331–349. https://doi.org/10.1007/BF01064504

Hedgecoe, A. M. (2004). Critical bioethics: beyond the social science critique of applied ethics. *Bioethics*, 18(2), 120–143. https://doi.org/10.1111/j.1467-8519.2004.00385.x

Hegel, G. W. F. (2017). *Phenomenology of Spirit*. Cambridge University Press. (First publ. 1807)

Holman, B., & Wilholt, T. (2022). The New Demarcation Problem. *Studies in History and Philosophy of Science*, 91, 211–220. https://doi.org/10.1016/j.shpsa.2021.11.011

Jaeggi, R. (2018): *Critique of forms of life.* Belknap Press.

Kahan, D. M., Dawson, E., Peters, E., & Slovic, P. (2013): Motivated Numeracy and Enlightened Self-Government. *Behavioural Public Policy*, 1(1), 54–86. https://doi.org/10.1017/bpp.2016.2

Kahneman, D., Slovic, P., & Tversky, A. (1982): *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press

Kant, I. (1998) (cited as CPR, A/B). *Critique of pure reason*. Cambridge University Press. (First publ. 1781)

Keren, A. (2013). Kitcher on Well-Ordered Science: Should Science Be Measured against the Outcomes of Ideal Democratic Deliberation?. *Theoria*, 28(2), 233–244.

Kitcher, P. (2011). *Science in a democratic society*. Prometheus Books.

Kourany, J. A. (2003). A Philosophy of Science for the Twenty-First Century. *Philosophy of Science*, 70(1), 1–14. https://doi.org/10.1086/367864

Kourany, J. A. (2008). Replacing the Ideal of Value-Free Science. In M. Carrier, D. Howard, & J. Kourany (Eds.), *The Challenge of the Social and the Pressure of Practice. Science and Values Revisited* (pp. 87–109). University of Pittsburgh Press.

Kroes, P., & Meijers, A. W. M. (2016). Toward an Axiological Turn in the Philosophy of Technology. In M. Franssen, P. E. Vermaas, P. Kroes, & A. W.M. Meijers (Eds.), *Philosophy of Technology after the Empirical Turn* (pp. 11–30). Springer International Publishing.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lewandowsky, S., & Oberauer, K. (2016): Motivated Rejection of Science. *Current Directions in Psychological Science*, 25(4), 217–222. https://doi.org/10.1177/0963721416654436

Lipsey, R. G., & Lancaster, K. (1956). The General Theory of Second Best. *The Review of Economic Studies*, 24(1), 11–32.

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Longino, H. E. (2004): How Values Can Be Good for Science. In P. Machamer, & G. Wolters (Eds.), *Science, Values, and Objectivity* (pp. 127–142). University of Pittsburgh Press.

Lövbrand, E., Pielke, R. A., & Beck, S. (2010). A Democracy Paradox in Studies of Science and Technology. *Science, Technology, & Human Values*. https://doi.org/10.1177/0162243910366154

McMullin, E. (1982). Values In Science. *Proceedings of the Biennial Meeting of the Philosophy of Science Association. Vol. Two: Symposia and Invited Papers*, 3–28. https://doi.org/10.1086/ psaprocbienmeetp.1982.2.192409

Millstone, E. (2005) Analysing the role of science in public policy-making. In P. van Zwanenberg, & E. Millstone, *BSE: Risk, Science and Governance* (pp. 11–38). Oxford University Press.

National Science Board (NSB) (2016): *Science and Engineering Indicators*. National Science Foundation. https://www.nsf.gov/statistics/2016/nsb20161/uploads/1/nsb20161.pdf

Popper, K. (1976). The Logic of the Social Sciences. In T. Adorno, H. Albert, R. Dahrendorf, J. Habermas, H. Pilot, & K. Popper (Eds.), *The Positivist Dispute in German Sociology* (pp. 87–104). Heinemann.

Rawls, J. (1955): Two Concepts of Rules. *The Philosophical Review*, 64(1), 3–32.

Reiners, D.S., Reiners, W.A., & Lockwood, J. A. (2013): The relationship between environmental advocacy, values, and science: a survey of ecological scientists' attitudes. *Ecological Applications*, 23(5), 1226–1242. https://doi.org/10.1890/12-1695.1

Reiss, J., & Sprenger, J. (2020). Scientific Objectivity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). https://plato.stanford.edu/entries/scientific-objectivity/

Robinson, K. S. (2015). *Aurora*. Orbit.

Rudner, R. (1953). The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science*, 20(1), 1–6. https://doi.org/10.1086/287231

Sanders, L. M. (1997): Against deliberation. *Political Theory*, 25(3), 347–376.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Searle, J. R. (1995). *The Construction of Social Reality*. The Free Press.

Simmons, A. J. (2010): Ideal and Nonideal Theory. *Philosophy & Public Affairs*, 38(1), 5–36. https:// doi.org/10.1111/j.1088-4963.2009.01172.x

Steel, B., List, P., Lach, D., & Shindler, B. (2004). The role of scientists in the environmental policy process: a case study from the American west. *Environmental Science & Policy*, 7(1), 1–13. https:// doi.org/10.1016/j.envsci.2003.10.004

Steel, D. (2016). Climate Change and Second-Order Uncertainty: Defending a Generalized, Normative, and Structural Argument from Inductive Risk. *Perspectives on Science*, 24(6), 696–721. https://doi.org/10.1162/POSC_a_00229

Steel, D., Gonnerman, C., McCright, A. M., & Bavli, I. (2018). Gender and Scientists' Views about the Value-Free Ideal. *Perspectives on Science*, 26(6), 619–657. https://doi.org/10.1162/posc_a_00292

Sunstein, C. R. (2006): Deliberating Groups versus Prediction Markets (or Hayek's Challenge to Habermas). *Episteme*, 3(3), 192–213. https://doi.org/10.3366/epi.2006.3.3.192

Valentini, L. (2012). Ideal vs. Non-ideal Theory: A Conceptual Map. *Philosophy Compass*, 7(9), 654–664. https://doi.org/10.1111/j.1747-9991.2012.00500.x

Van der Hel, S. (2018). Science for change: A survey on the normative and political dimensions of global sustainability research. *Global Environmental Change*, 52, 248–258. https://doi.org/10.1016/ j.gloenvcha.2018.07.005

Weber, M. (1949). "Objectivity" in Social Science and Social Policy. In *On the Methodology of the Social Sciences* (pp. 50–112). The Free Press. (First publ. 1904)

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science*, 40(1), 92–101. https://doi.org/10.1016/j.shpsa.2008.12.005

Wilholt, T. (2014). Philip Kitcher, Science in a Democratic Society. *Philosophy of Science,* 81(1), 165-171. https://doi.org/doi:10.1086/674367

Winsberg, E. (2012). Values and Uncertainties in the Prediction of Global Climate Models. *Kennedy Institute of Ethics Journal*, 22(2), 111–137. https://doi.org/10.1353/ken.2012.0008

Young, I. M. (2001). Activist Challenges to Deliberative Democracy. *Political Theory*, 29(5), 670–690. https://doi.org/10.1177/0090591701029005004

V. MODELS OF SCIENCE AND SOCIETY: TRANSCENDING THE ANTAGONISM*

ZUSAMMENFASSUNG Das im vierten Teil diskutierte Thema der Stakeholderbeteiligung wirft die Frage auf, welches Verhältnis von Wissenschaft und Gesellschaft grundsätzlich angemessen ist. Im fünften Teil betrachte ich diese Frage aus einer übergeordneten Perspektive. Ich diskutiere, wie konkurrierende Antworten auf die Frage systematisiert und konstruktiv aufeinander bezogen werden können. Grundlage hierfür ist das Konzept der *Wissenschafts-Gesellschafts-Interaktionsmodelle*. Anders als die vorangegangenen Teile ist der Teil interdisziplinär und anwendungsorientiert. Weiterhin bezieht er sich auf eine vor allem in den Science and Technology Studies und der transdisziplinären Forschung geführte Diskussion. Ich kritisiere die in diesen Kontexten verbreitete Tendenz, Interaktionsmodelle als widerstreitende Lager oder Repräsentationen von Akteursüberzeugungen zu interpretieren. Diese Interpretation geht häufig mit einer strikten, mitunter polemisch vorgetragenen Ablehnung bestimmter Interaktionsmodelle einher. Ich zeige dies anhand der populären Unterscheidung zwischen *technokratischen*, *dezisionistischen* und *pragmatistischen* Modellen. Ich argumentiere, dass diese Modelle sinnvoll koexistieren können, wenn sie als Idealtypen und Heuristiken verstanden werden. Auf dieser Basis erarbeite ich Vorschläge für ein *Reflexionstool*, das in realen Wissenschafts-Gesellschafts-Interaktionen genutzt werden kann, um Hintergrundannahmen über Wertfreiheit und andere Grundsatzfragen explizit zu machen und produktiv zu diskutieren.

* Dieser Artikel wurde zuerst veröffentlicht als Dressel, M. (2022). Models of Science and Society: Transcending the Antagonism. *Humanities and Social Sciences Communications*, 9, 241. https://doi.org/10. 1057/s41599-022-01261-x. Teile dieses Artikels sind außerdem veröffentlicht in Dressel, M. (2022). *Wissenschaft und Gesellschaft. Modelle des Wissenschafts-Gesellschafts-Verhältnisses und ihre Hintergrundannahmen*. Climate Service Center Germany.

1. Introduction

What is the appropriate place of science in society? Despite the vast literature on the subject, the science-society relation remains a highly disputed subject (Lincoln & Guba, 2000; Millstone, 2005; Hulme, 2009; Skelton, 2021). This is not only due to the many contexts in which scientists and non-scientists interact, e.g. in policy advice (Kowarsch, 2016), science communication (Brossard & Lewenstein, 2010), or stakeholder encounters within research processes (Bremer & Meisch, 2017). Determining an appropriate role for science in society is also difficult because the question itself is overly complex: Should science be autonomous (Wilholt, 2010; Kitcher, 2011)? Can and should research processes be free of social values (Lacey, 1999; Longino, 2002; Douglas, 2009)? Should scientists advocate societal change (Nielsen, 2001; Pielke, 2012)? Does rational action presuppose scientific evidence (Oreskes, 2004; Sarewitz, 2004)? Hence, when asking about the right place for science in society, we are actually asking various interrelated, non-trivial individual questions. Many of these are "deep-seated, normative questions" (Miller, 2001, p. 479), and most of them are "fiercely contested" (Millstone, 2005, p. 11).

A sensible way to organize this complexity are science-society interaction models (SSIMs). By aggregating the manifold aspects of the science-society relation into clear-cut concepts, SSIMs provide generic, easy to grasp templates for interactions between scientists and non-scientists. This can be helpful both for scholars who study the science-society relation and for real-world actors (scientists and non-scientists) who engage in science-society interactions. However, as science's ideal role in society is contested, a range of SSIMs is conceivable. Scholars have therefore proposed various SSIM taxonomies, i.e. sets of SSIMs that describe opposing views of the science-society relation (e.g. Gibbons et al., 1994; Lincoln & Guba, 2000; Jasanoff, 2003; Trench, 2008; Hessels et al., 2009; Pielke, 2012; Krishna, 2014; Fazey et al., 2018; Skelton, 2021). One prominent example is the distinction between technocratic, decisionist and pragmatist models (TDP taxonomy) (Habermas, 1970). While there are many SSIM taxonomies on the market, the TDP taxonomy remains a popular analytical lens, particularly regarding science-society interactions in policy contexts (Weingart, 1999; Brown et al., 2005; Heinrichs, 2005; Millstone, 2005; Lompe, 2006; Hulme, 2009; Gluckman, 2011; Edenhofer & Seyboth, 2013; Kowarsch, 2016).

This paper discusses the benefits and limitations of SSIMs, focusing on the TDP taxonomy. The paper pursues two aims: first, it scrutinizes how SSIMs should be understood

from a theoretical perspective; second, it discusses how SSIMs may be of practical use for actors who wish to reflect on their fundamental (or "philosophical") science-society assumptions. The paper argues that SSIMs are valuable for both theoretical and practical purposes. At the same time, however, there is a tendency in the science-society literature that undermines these merits, namely the tendency to interpret SSIMs as antagonistic theoretical camps or as representations of the beliefs and attitudes of real-world actors. The paper argues that these pitfalls can be avoided if we take SSIMs for what they really are: nothing more – and nothing less – than ideal-types and heuristics. While similar concerns have been voiced before, both in the context of the science-society relation (Trench, 2008; Martin, 2012; Jahn et al., 2021) and in the context of the philosophy of modeling (Morgan & Morrison, 1999; Giere, 2004; Colyvan, 2013), this view of SSIMs has not been systematically explored in the science-society literature so far. Building on this interpretation, the paper presents some preliminary steps towards a reflexive tool that actors may use to unwrap their implicit assumptions about science and society. The tool involves six dimensions of key questions and a modified version of the TDP taxonomy. The combination of these dimensions and the modified TDP taxonomy describes a conceptual space in which actors can identify, compare and discuss their science-society assumptions in a constructive manner. The results of such a discussion may be used to design, e.g., research projects or advisory processes at the science-society interface.

2. SSIMs: definition and examples

2.1 WHAT IS AN SSIM?

This section discusses exemplary SSIMs from the literature, with special consideration of the TDP taxonomy. For this purpose, we first need to define the term SSIM:

Def. 1 A science-society interaction model (SSIM) represents the totality of fundamental assumptions that an actor holds regarding the way in which scientists and non-scientists should interact in a given context. SSIMs involve deep-rooted ("philosophical") assumptions about the nature of scientific inquiry and the principles of social order. These assumptions may be explicit or implicit, stronger or weaker, stable or flexible. Relevant contexts include, *inter alia*, policy advice,

> science communication, or stakeholder interactions within research processes. Actors may be individuals (scientists, regulators etc.) or collective entities (institutions, associations etc.).

While SSIMs have been extensively discussed by scholars of science and society (see sect. 2.2), this actor-specific interpretation suggests that an SSIM need not necessarily have an equivalent in the academic literature. That is, a real-world actor can hold certain beliefs on science and society that are, in this particular configuration, not at all discussed in the relevant literature; yet, the actor's set of assumptions is still an SSIM, however implicit or theoretically unarticulated these assumptions may be. SSIMs that are part of a published taxonomy are termed *taxonomic* in this paper, while SSIMs that real-world actors (often implicitly) hold are termed *non-taxonomic*. Also, note that the above definition refers to normative SSIMs. This does not mean that SSIMs do not include descriptive assumptions; however, the SSIMs considered in this paper are normative in that they provide competing answers to the question how scientists and non-scientists *should* interact. While this paper does not take any position as to which of these answers is ultimately correct, it suggests that it is worthwhile to make the underlying science-society assumptions explicit and discuss them in a constructive manner (see sect. 4.3). Furthermore, this paper focusses on fundamental or "philosophical" (in a broad sense of the word) aspects of SSIMs, acknowledging that actors will often hold these philosophical assumptions in an implicit rather than explicit manner (Hulme, 2009, p. 94; Kowarsch 2016, p. 82; Crowley & O'Rourke, 2021).

Finally, it is important to avoid two misunderstandings. One misunderstanding would be that terms such as "science-society relation" or "science-society interaction" are meant in a literal or reifying sense. This would be implausible, not only because science is obviously part of society, but also because science and society cannot interact as a whole (only specific actors can do so). Also, there are of course many global societies and many scientific disciplines. Terms such as "science-society relation" or "science-society interaction" should thus not be taken literally, but as a linguistic convention that captures a large variety of interactions between scientists and non-scientists in one simple term.

Another misunderstanding would be that SSIMs are "models" in the same sense as, e.g., the Bohr model of the atom or a general circulation model of the global climate. While these types of models resemble SSIMs by providing stylized representations of a target system (see e.g. Frigg & Hartmann, 2020) and by serving some of the cognitive functions that SSIMs serve (e.g. complexity reduction), they differ from SSIMs in several ways.

First, SSIMs are normative rather than descriptive. In this respect, SSIMs are more similar to ideals in moral or political philosophy than to models used in the natural or social sciences. Second, SSIMs are neither predictive, nor do they allow for causal inferences. SSIMs can therefore not be "fed with data" to draw conclusions about how a target system may behave under similar conditions. Third, SSIMs are sets of assumptions, but these assumptions are not linked in the same sense in which, say, atmospheric and ocean components of a climate model are coupled. Rather, SSIMs form a *semantic web*, i.e. a network of beliefs that share a conceptual affinity with each other (similar to the affinity that Max Weber famously saw between Protestantism and capitalism). Fourth, SSIMs are not even consistently called "models" in the science-society literature, with terms such as "paradigms" (Lincoln & Guba, 2000), "modes" (Gibbons et al., 1994), "social contracts" (Hessels et al., 2009) or "orders" (Skelton, 2021) offering alternatives that do not evoke the association with scientific models. Despite the similarities that exist between SSIMs and scientific models, it thus seems plausible to treat SSIMs as an own ontological and methodological category.

2.2 TAXONOMIC SSIMS: EXAMPLES

The question of how science and society should interact has been discussed many times in the academic literature (e.g. Habermas, 1970; Jasanoff, 1990; Gibbons et al., 1994; Guston, 2001; Kourany, 2003; Longino, 2002; Estlund, 2003; Nowotny, 2003; Sarewitz, 2004; Carrier et al., 2008; Brown, 2009; Douglas, 2009; Hoyningen-Huene, 2009; Elliott, 2011; Kitcher, 2011; Edenhofer & Kowarsch, 2015; Bremer & Meisch, 2017). Apart from certain trends, however, the question remains "a fiercely contested domain" (Millstone, 2005, p. 11). Scholars disagree about fundamental issues such as scientific value-freedom (e.g. Douglas, 2009 vs. Betz, 2013) or political neutrality (e.g. Nielsen, 2001 vs. Hagedorn et al., 2019); also, the literature is divided into various disciplines, theoretical approaches and terminologies. Moreover, the debate is fragmented into multiple thematic contexts such as policy advice (e.g. Jasanoff, 2003; Pielke, 2012), science communication (e.g. Trench, 2008; Brossard & Lewenstein, 2010) or analyses of research paradigms (e.g. Fazey et al., 2018; Skelton, 2021). The "substantial literature" (Millstone, 2005, p. 11) on the science-society relation thus "does not provide anything resembling a single, coherent agreed framework" (ibid.).

Yet despite the differences, the literature features a communality: the use of conceptual oppositions. Popular examples include "tame" versus "wicked problems" (Rittel & Webber, 1973), "issue advocacy" versus "honest brokering" (Pielke, 2012), "normal" versus

"post-normal science" (Funtowicz & Ravetz, 1993), and many others. The idea of these concept pairs is that there are fundamentally different ways of doing science, and that the question of how science and society should relate to each other can be answered in very different ways. Of course, not all concept pairs in the science-society literature qualify as full-fledged SSIMs. The famous tame-wicked distinction, for instance, addresses the more specific question how socio-scientific problems should be understood (see also Schmidt, 2011). While this is not *as such* an SSIM in the above sense, such distinctions have implications for the science-society relation at large, e.g. whether or not the public should "entrust *de facto* decision-making to the wise and knowledgeable professional experts" (Rittel & Webber, 1973, p. 169). Also, these concept pairs presuppose assumptions of a very general kind, e.g. whether there are "value-free, true-false answers" (ibid.) in science, or whether society is characterized by a "politicization of subpublics" (ibid, p. 167). These assumptions reach far beyond seemingly narrow questions such as "what is a socio-scientific problem?" (Schmidt, 2011) and touch upon many dimensions that a full-fledged SSIM includes. We can thus take it that, while not all concepts in the science-society literature are SSIMs as such, they have strong conceptual ties with the larger SSIM taxonomies that we discuss in this paper.

A crucial aspect, and a starting point of this paper, is that SSIMs have similar structures and recurring themes. Consider the following examples from different branches of the science-society literature. In the context of the science communication, authors have contrasted the "deficit model" with concepts such as the "interactive science model" (Einsiedel, 2000). The former claims that "the public [is] 'deficient', while science is 'sufficient'" regarding knowledge quality (Sturgis & Allum, 2004, p. 57). In contrast, the latter model emphasizes "the uncertainty of scientific knowledge" (Einsiedel, 2000, p. 144) or "the lack of demarcation between scientific […] and other forms of knowledge" (ibid.). In the context of research paradigms, Lincoln & Guba (2000) make a similar distinction between "old" and "new paradigm research". The former claims that science should be objective and detached from social practice, whereas the latter holds that "social transformation […] is the end goal [of science]" (ibid., p. 172), and that research is "incomplete without action on the part of participants" (ibid.). In a similar context, Gibbons et al. (1994) have famously distinguished "mode 1" from "mode 2" research. The former is "characterized by the hegemony of theoretical or […] experimental science; by an internally-driven taxonomy of disciplines; and by the autonomy of scientists" (Nowotny et al., 2003, p. 179). The latter is "socially distributed, application-oriented, trans-disciplinary, and subject to multiple accountabilities" (ibid.). As we shall see in the next subsection, the sort of assumptions that these distinctions refer to, but also the way

in which the assumptions are contrasted with each other, are typical for SSIM taxonomies in the science-society literature. While this paper focusses on the TDP taxonomy, a good part of the following considerations can thus be transferred to other taxonomies and concept pairs in the field.

2.3 THE TDP TAXONOMY

Let us now discuss the technocratic-decisionist-pragmatist (TDP) taxonomy. Originally proposed by Habermas (1970), the taxonomy has become a "traditional" (Lompe, 2006) analytical lens for the relation between science and societal decision-making, particularly in public policy (Weingart, 1999; Brown et al., 2005; Heinrichs, 2005; Millstone, 2005; Hulme, 2009; Gluckman, 2011; Edenhofer & Seyboth, 2013; Kowarsch, 2016). The taxonomy is also the basis for the widely discussed "linear" model of expertise, which is largely identical to the technocratic (Beck, 2011) or the decisionist model (Heazle et al., 2016), depending on the interpretation (Durant, 2016). It has also been used as an umbrella concept for both of these models (Weingart, 1999; Lompe, 2006). There are three reasons why this paper focusses on the TDP taxonomy rather than, say, the mode 1/2 distinction. First, the TDP taxonomy is not only one of the most popular SSIM schemes in the literature, it also looks back on a longer history than other taxonomies. The fact that after more than fifty years Habermas' SSIMs are still "mentioned again and again" (Grunwald & Saretzki, 2019, pp. 12-13, my transl.) makes this taxonomy particularly interesting. Second, the TDP taxonomy is more generic than many dualist taxonomies ("mode 1" vs. "mode 2", "normal" vs. "post-normal science", "old paradigm" vs. "new paradigm research" etc.). One advantage of the TDP taxonomy is therefore that, as Martin Kowarsch has put it, various other SSIMs "can be understood as mere variations or mixtures of the three models presented by Habermas" (2016, p. 85). Third, the TDP taxonomy can be extended beyond its original context, scientific policy advice, such that it includes a range of other aspects of the science-society relation (e.g. science communication or co-productive research, see sect. 4.2). Hence, although the TDP taxonomy is but one of many SSIM schemes on the market, it is particularly promising as a starting point for analyzing the science-society relation. The TDP taxonomy comprises three SSIMs[1]:

1 Several authors have proposed modifications of the classic TDP taxonomy, leading to a greater (see e.g. Millstone, 2005; Kowarsch, 2016) or a smaller number of taxonomy members (e.g. when technocratic and decisionist models are subsumed under the linear model, see Weingart, 1999; Lompe, 2006; Durant, 2016). A variation of the decisionist model was already discussed in Habermas (1970). But contrary to the classic version of the TDP taxonomy, these modifications have either not been widely adopted, or are not used in a consistent sense throughout the literature.

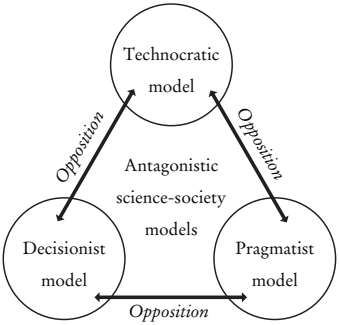
(T) The technocratic model aims to rationalize society by subverting it to the "objective knowledge of the expert" (Habermas, 1970, p. 63). This includes both the determination of means (technologies, strategies) and ends (practical goals). Science determines means and ends in a value-free manner (Kowarsch, 2016, p. 88). The SSIM hence suggests that "technical considerations are not just necessary but also sufficient for policy decision-making" (Millstone, 2005, p. 14).

(D) The decisionist model aims "to separate strictly the functions of the expert from those of the politician [and the decision-maker more general]" (Habermas, 1970, p. 63). However, this only holds for societal ends, which the SSIM assumes to be value-laden. But once the goals are set by societal actors, the means can and should be determined objectively by science. The SSIM's main features are thus the value-freedom of the research process and a neutral role for science in societal debates (Millstone, 2005; Kowarsch, 2016).

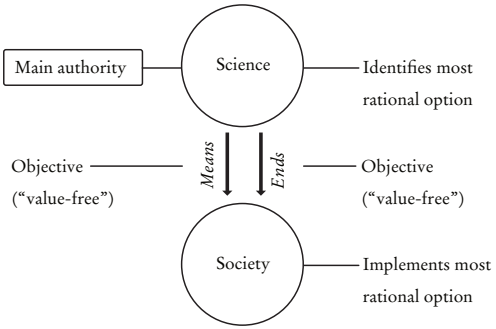
(P) The pragmatist model envisions an iterative process, where science "is governed by a horizon of […] value systems" (Habermas, 1970, p. 67), and where social values are "being tested with regard to the technical possibilities [as identified by science]" (ibid.). While science actively shapes societal goals, it has no unquestionable authority. Rather, fact finding and norm setting are interdependent (Edenhofer & Kowarsch, 2015). The SSIM thus rejects both value-free research and the neutrality of science in public debates. The pragmatist model, or versions of it, represent the dominant trend in the current science-society literature (Kowarsch, 2016, p. 91).

As can be seen from these descriptions, the TDP taxonomy addresses similar issues as the distinctions mentioned in the previous section, e.g. objectivity, authority, or neutrality. Even more crucially, these taxonomies have similar structures: First, they typically involve only few SSIMs (three in the case of the TDP taxonomy). Second, taxonomies describe SSIMs in a rough, schematic way. Some of these schematic descriptions imply strong or even extreme positions (e.g. "scientists should have full decision authority in society"). Third, taxonomies typically assume sharp oppositions between SSIMs, with clear-cut conceptual divides and contradictory views of science and society. This incentivizes a view that will be critically discussed in the next section: the idea that SSIMs describe opposing theoretical camps, and that an actor who subscribes to one taxonomic SSIM will (or should) reject the other taxonomy members. An illustration of the TDP taxonomy and the respective SSIMs is given in figure 1.

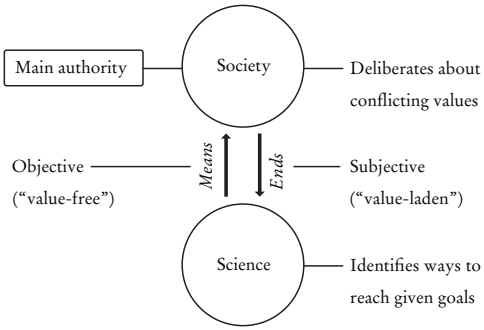
(1a) The TDP taxonomy



(1b) The technocratic model (T)



(1c) The decisionist model (D)



(1d) The pragmatist model (P)

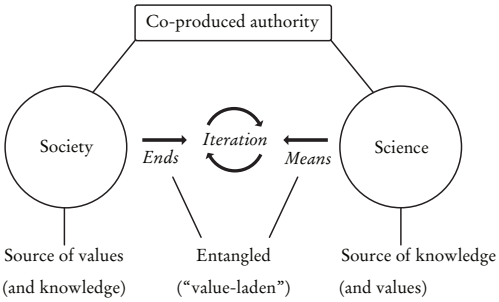


Figure 1: The TDP taxonomy. The included SSIMs are typically conceived as antagonistic (a). The technocratic model assigns a strong decision authority to science (b). The decisionist model assigns the main decision authority to society (c). In the pragmatist model, the decision authority is co-produced in an iterative exchange between science and society (d).

3. The problem with taxonomic SSIMs

3.1 TAXONOMIC SSIMS: BENEFITS

This section discusses the benefits and limitations of SSIMs and SSIM taxonomies. It argues that SSIMs can be valuable for both theoretical and practical purposes. However, they also have downsides, at least in a certain interpretation. The section argues that in order to transcend the limitations while keeping the benefits, we should strictly treat taxonomic SSIMs as ideal-types and heuristics. While this interpretation is not new in itself, it is obscured by a tendency in the literature to view taxonomic SSIMs as antagonistic camps or as representations of the beliefs and attitudes of real-world actors.

SSIMs and SSIM taxonomies have two basic merits: they reduce complexity, and they illuminate the contestedness of the science-society relation. The first point refers to the complicated character of the science-society relation. For instance, whether science should be autonomous, whether science is superior to other types of knowledge, and what this means for, say, the role of scientists in public debates – these are non-trivial questions. SSIMs organize this complexity (in this regard they are similar to models in other contexts, see e.g. Turnbull, 1993; Frigg & Hartmann, 2020). By reducing the range of conceivable views to a manageable number of easy to grasp concepts, they provide analytical templates by which scholars can categorize different claims and theories about the science-society relation; real-world actors (scientists and non-scientists), on the other hand, may find it easier to judge the quality of a science-society interaction if they know which options are theoretically conceivable. The most crucial benefit of SSIMs is thus their orienting cognitive function.

The second benefit refers to the controversial character of the science-society relation. Debates about the right place for science in society reach back centuries (Hessels et al., 2009; Martin, 2012; Krishna, 2014) and are unlikely to be settled any time soon. Empirical research also shows that actors disagree on these issues. For instance, Van der Hel (2018) finds that "[s]ome researchers perceive it as not only impossible but also undesirable to separate normative and value-laden questions from […] research", whereas other scientists hold "that researchers should strive for independence and impartiality" (ibid., p. 256). Discrepancies can also be found between actor groups. Steel et al. (2004), for instance, find that only 16% of researchers in their sample agree that scientists should advocate for specific policies; representatives of interest groups and the public, however, are much more likely to agree with that claim (46% and 36%, respectively, ibid., pp. 6–7). Furthermore, these findings vary strongly across studies (see e.g. Gray & Campbell, 2008; Reiners et al., 2013). SSIM taxonomies make these controversies visible. This helps scholars to identify fundamental differences between competing theories, and may support real-world actors in understanding the extent and gravity of potential disagreements in a science-society interaction.

3.2 TAXONOMIC SSIMS: WEAKNESSES

Yet these benefits come at a price. As indicated before, taxonomies of SSIMs tend to be narrow (involving only few conceptual options), stereotypical (involving undifferentiated and sometimes extreme positions), and antagonistic (involving harsh conceptual oppositions). While these are exactly the properties that make SSIM taxonomies valuable

as reducers of complexity and illuminators of conceptual divides, they can also give rise to a limited and antithetical view of SSIMs. Consider the issues of authority, objectivity and neutrality. The TDP taxonomy presents us with the following alternative:

(T) Scientists should have decision authority, as they possess the objectively correct solutions to societal problems.
(D) Scientists should remain neutral, as societal goals are normative and inaccessible to objective scientific analysis.
(P) Scientists should neither have decision authority, nor can they remain objective and neutral, as their expertise is deeply value-laden.

We can easily see that the spectrum described by these claims is not very nuanced. Moreover, these claims can create the impression that accepting one SSIM forces us to reject the other two. It may be objected that scholars are well aware of the many middle ground positions that are possible between these extremes, and that consequently SSIMs are not seen as mutually exclusive in the literature. Unfortunately, however, many scholars have embraced exactly this antagonistic interpretation. As Brossard & Lewenstein (2010) have put it: "The literature tends to describe these models as mutually exclusive and to present them as the backbone of different research and outreach paradigms" (ibid., p. 17). For instance, Lincoln & Guba (2000) argue in their influential article on paradigmatic science-society controversies that SSIMs are "incommensurable" (ibid., p. 172), and that actors cannot "'pick and choose' among the axioms of [opposing] models, because the axioms are contradictory and mutually exclusive" (ibid., p. 174). While Lincoln & Guba concede that there are overlaps between those SSIMs that are part of one SSIM family, they argue that it is "highly unlikely and would probably even be less than useful" (ibid., p. 185) that proponents of opposing SSIMs try to "find some way of resolving their differences" (ibid.). As a direct consequence of this antagonistic view, many scholars believe that one SSIM – typically a version of the pragmatist model (Pielke, 2012: 78; Durant, 2016; Kowarsch, 2016, p. 91) – should straightforwardly replace its competitors. In this vein, critics of the linear model(s) have made it very clear that their aim is not to identify middle grounds and overlaps between SSIMs, but to substitute one SSIM with another. Only very rarely have scholars argued that the technocratic and decisionist models should co-exist with the pragmatist model (a notable exception is Durant, 2016); rather, the dominant idea is that the technocratic and decisionist models are "significantly flawed" (Beck, 2011, p. 304), "profoundly mistaken" (Grundmann & Rödder, 2019, p. 9), weak (Habermas 1970, p. 64) or dubious (Habermas 1970, p. 65), and that consequently these SSIMs should be abandoned once and for all.

The problem with this view is not that the often criticized decisionist and technocratic (or linear) models were actually "true" – in fact, it may not even be relevant whether a taxonomic SSIM is "true" (see sect. 3.4). The problem is rather the underlying idea that subscribing to one taxonomic SSIM means rejecting the other two. Yet in reality, there are many positions that overlap with *all* of the above claims, and these positions are neither internally inconsistent, nor "significantly flawed" or "profoundly mistaken" despite their partial acceptance of claims (T) and (D). For instance, an actor who believes that value-ladenness is inevitable may still hold that values are less relevant in some branches of research than in others. Similarly, an actor may agree that neutrality is not fully achievable in societal debates, but still maintain that some styles of scientific advice are more neutral, and thus more legitimate, than others. An actor may also believe that science should not hold decision authority in general, and yet hold that technocratic decisions are more legitimate in some cases than in others. These alternatives imply a partial and conditional acceptance of all three of the above claims – obviously without any formal or material contradiction. It is thus misleading to interpret taxonomic SSIMs as opposing theoretical camps that cannot co-exist in a reasonable and consistent way.

Another problem is that the above claims seem to suggest a straightforward link between neutrality, objectivity, and value-ladenness. Yet in reality, a range of positions is conceivable here. For instance, some authors hold that science can be value-laden and yet objective (Harding, 1995; Steel et al., 2017), while others hold that *if* science is value-laden, *then* "[t]he research process can no longer be characterized as an 'objective' investigation" (Nowotny et al., 2003, p. 187). Similarly, some hold that *if* science is value-laden, *then* scientific advice cannot be neutral (Betz, 2013), while others argue that neutrality is a function of high epistemic standards rather than value-freedom (John, 2015). This reinforces the point that taxonomic SSIMs are not mutually exclusive, and that the focus on the conceptual differences between SSIMs can obscure the large spectrum of conceivable middle ground positions.

It may be objected that these issues are merely theoretical, as taxonomic SSIMs are not meant to guide real-world science-society interactions. But this is not how SSIMs are discussed in the literature. Quite on the contrary, many authors have argued that SSIMs have an "action-guiding character" (Kowarsch, 2016, p. 83) (see also Hulme, 2009, ch. 3.5; Hubbs et al., 2021). Mike Hulme has famously argued that "part of the reason that we disagree about climate change is that there are a number of different models of how science is (or should be) used in policy development" (2009, p. 94) and that "recognising these different models of science-policy interaction is crucial" (ibid.) for both scientists

and policy-makers. A dominant motivation in the literature is therefore to advance science-society interactions by changing the underlying SSIMs (e.g. Habermas, 1970; Jasanoff, 2003; Sarewitz, 2004; Beck, 2011; Pielke, 2012; Grundmann & Rödder, 2019). It is hence safe to assume that academic debates about SSIMs are not purely theoretical, but also aim to assist real-world actors (for a practical example, see Felt et al., 2007).

The problem, however, is that the antagonistic interpretation of the TDP taxonomy is not well suited for this purpose. Actors may not find it convincing that SSIMs are "mutually exclusive" (Lincoln & Guba, 2000, p. 174), or that the decisionist and technocratic models are so "profoundly mistaken" (Grundmann & Rödder, 2019, p. 9) that there cannot be a reasonable compromise that also includes some decisionist or technocratic elements. While empirical studies find correlations between actors' assumptions regarding different aspects of science and society (Gray & Campbell, 2008; Reiners et al., 2013; Steel et al., 2017), they do not support the idea that actors are strict supporters of one single taxonomic SSIM. For instance, an actor who believes that science is the only reliable form of knowledge can – but need not necessarily – believe that scientists should make societal decisions (Steel et al., 2004, p. 7). Empirically speaking, actors are typically *somewhat* technocratic in *some* respects and *somewhat* decisionist or pragmatist in *other* respects. This, however, is not reflected by the antagonistic interpretation of the TDP taxonomy.

3.3 IS THE CRITIQUE OF TAXONOMIC SSIMS OLD NEWS?

At first glance, the critique that there are many middle ground positions between taxonomic SSIMs may seem like a truism. Also, one might wonder whether similar critiques that have long been discussed in other contexts, e.g. in scientific modelling (Box, 1979) or regarding idealizations more generally (Turnbull, 1993), have not been considered in debates about the science-society relation. However, several considerations indicate that the apparent truism is not so trivial after all. The first consideration is that, if the critique were indeed common sense in the science-society literature, one would not expect to find concerns regarding the antagonistic interpretation in this literature. Yet several authors have raised exactly this concern. Sturgis & Allum (2004) contend that scholars have focused too much on extreme versions of the deficit model of science communication, and that this taxonomic SSIM is "something of a 'straw man'" (ibid., p. 57); Trench (2008) argues that the literature has adopted a "bipolar view" (ibid., p. 130) of SSIMs, and that this view is "neither an accurate account of recent developments nor a useful guide to current and future practice and analysis" (ibid.); Jahn et al. (2021) maintain that

the literature "seems to imply a duality of transdisciplinary [i.e. stakeholder-involving] versus non-transdisciplinary research" (ibid., p. 2), and that scholars lost sight of the "spectrum of more or less transdisciplinary research modes" (ibid.) that we find in reality; Martin (2012) argues that the famous mode 1/2 distinction is often interpreted as mutually exclusive, and that in reality these SSIMs come in mixtures. These remarks indicate that the antagonistic interpretation is indeed widespread; however, one should also note that such remarks are scattered in the literature and have not been systematically explored on a general level.

Another consideration is that, if the above critique were indeed a truism, one would expect that the technocratic and the decisionist models are discussed in such a way that these SSIMs are at least partially acceptable. Also, one would expect scholars to emphasize that there is not typically a full convergence between an actor's assumptions and a taxonomic SSIM. Yet these expectations prove to be false. Rather, the terms "decisionist", "technocratic" and "linear" have long been vehicles to criticize scientists for being ignorant of their own normativity and the nature of politics (starting with Habermas, 1970). Furthermore, scholars have often portrayed actors as exclusive subscribers to one single SSIM. Beck (2011), for instance, has argued that the Intergovernmental Panel on Climate Change (IPCC) "clearly uses the linear model of expertise" (ibid., p. 298), and Grundmann & Rödder (2019) have argued that "key participants in the climate change discourse operate under the assumption of a linear model" (ibid., p.1). The idea that aspects of the linear model may be worth keeping, or that actors' assumptions are more nuanced than the schematic description of a taxonomic SSIM, is not very typical for these contributions.

This brings us to a final consideration. If the discussed critique were a truism, it would be implausible for scholars to anticipate that an SSIM vanishes completely from science-society debates. After all, it seems perfectly acceptable that taxonomic SSIMs co-exist if they are not antagonistic and mutually exclusive. Yet there are many complaints in the literature that, e.g., the linear model is "refusing to die" (Durant, 2016, p. 31) despite scholars of science and society "dealing the linear model of expertise innumerable mortal blows" (ibid., p. 17). A sensible explanation why many scholars find it "puzzling" (Van der Hel, 2018, p. 256) that this SSIM "remains to receive such strong support" (ibid.) is that they assume the antagonistic interpretation discussed above. The apparent truism that actors may well support several taxonomic SSIMs at the same time is therefore not so trivial after all.

3.4 REVISITING TAXONOMIC SSIMS

The previous subsections have argued that, on the one hand, SSIMs and SSIM taxonomies can be useful for theoretical and practical purposes; on the other hand, however, they can give rise to an unconstructive view of the respective debates. This raises the question how taxonomic SSIMs can be used in a more constructive manner, thus overcoming the shortcomings while keeping the benefits. Building on the above considerations, a revised understanding should have the following characteristics:

* *Wider option space.* A revised understanding should allow for more SSIM alternatives, while also reducing complexity in such a way that taxonomic SSIMs retain their orienting function.
* *Non-stereotypical SSIMs.* A revised understanding should avoid the impression that actors need to commit to a taxonomic SSIM in its purest (most "extreme") form.
* *Non-antagonistic SSIMs.* A revised understanding should show how actors can partially accept several taxonomic SSIMs simultaneously, while also illuminating the divides between SSIMs.

This paper argues that, in order to achieve these goals, we should refrain from treating taxonomic SSIMs as theoretical camps or representations of actor beliefs. Rather, we should take seriously the remarks made by several authors that taxonomic SSIMs are no more – and no less – than *ideal-types* and *heuristics* (Heinrichs, 2005; Lompe, 2006; Trench, 2008; Kowarsch, 2016, p. 82-85; Jahn et al., 2021). In their capacity as ideal-types, taxonomic SSIMs are stylized constructions that illustrate how the science-society relation would look like if the assumptions of the SSIMs would manifest themselves in a pronounced or even extreme form (Weber, 1904/1949). In their capacity as heuristics, taxonomic SSIMs are landmarks that help scholars and science-society actors navigate the landscape of possible positions.

Importantly, the ideal-typical and heuristic interpretation is only applicable to taxonomic SSIMs. The set of non-taxonomic SSIMs, on the other hand, should do justice to the large spectrum of "non-strawman" assumptions that real-world actors (scientists and non-scientists) actually hold. A core claim of this paper is that *taxonomic SSIMs can be used to describe a conceptual space in which we can identify, compare and critically assess real-world assumptions in a constructive manner.* Taxonomic SSIMs promote such a constructive discussion by providing us with boundary cases. Real-world actors' assumptions may approximate these boundary cases, but will not typically be identical

with them. Also, actors' assumptions can, and typically will, be pulled towards different taxonomic SSIMs at the same time. This is because actor assumptions are not monolithic, but complex *semantic webs*, i.e. networks of beliefs that actors hold regarding different aspects of science and society (for an illustration, see figure 2).

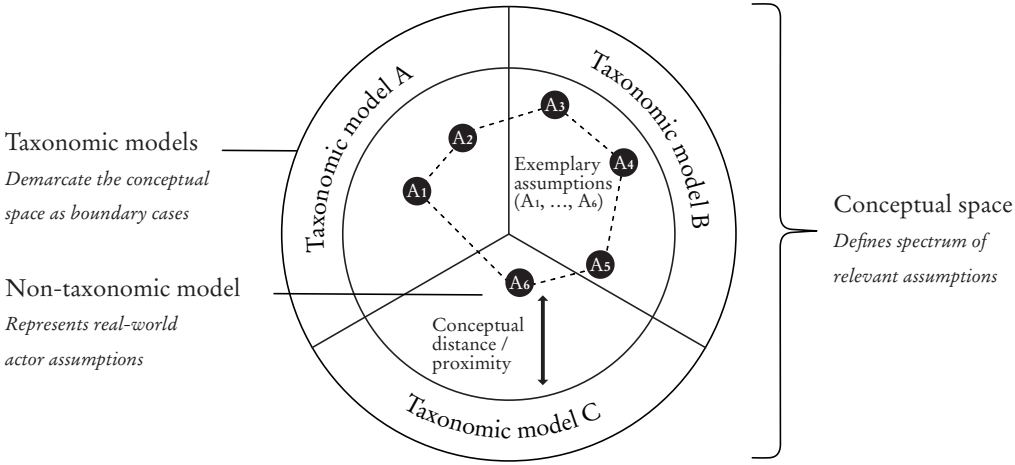


Figure 2: The conceptual space. Taxonomic SSIMs are used as boundary cases. Real-world assumptions are defined by their conceptual distance/proximity to the boundary cases. Real-world actors will typically hold several relevant assumptions. The totality of an actor's assumptions makes up the actor's non-taxonomic SSIM.

4. Application: tentative steps towards a reflexive tool

4.1 SIX DIMENSIONS OF THE SCIENCE-SOCIETY RELATION

While the previous section proposed a theoretical understanding of SSIMs, the following section discusses its practical application. The section rests on an idea mentioned before: that SSIMs and SSIM taxonomies are valuable not only for scholars who study the science-society relation, but also for actors (scientists and non-scientists) who engage in real-world science-society interactions. The section presents a first, preliminary sketch of a reflexive tool that may be used to identify and discuss the underlying assumptions of such interactions. The results may be used to design, e.g., advisory processes or research projects. The section discusses three steps, each of which can be adapted to the specifics of a given science-society interaction: First, the identification of fundamental or "philosophical" questions that are relevant to the interaction at hand (this subsection); second, the construction of taxonomic SSIMs that provide ideal-typical answers to these questions (sect. 4.2); and third, the choice of a method for mapping the actors' assumptions in the resulting conceptual space (sect. 4.3).

Starting with the first step, it is useful to discern those aspects of a science-society interaction that touch upon the nature of scientific inquiry (science-related dimensions) from those aspects that focus on the principles of social order and human behavior (society-related dimensions). On a more concrete level, this paper suggests to differentiate between the dimensions epistemic standards, epistemic scope, epistemic interests, social legitimacy, social rationalities and social structure. These dimensions provide categories under which a variety of key questions can be subsumed. The key questions are more specific than the six dimensions and may be weighted differently in different contexts. Also, actors may identify questions that are unique to a given interaction. However, starting from the literature[2] we can determine a tentative list of questions that are generally relevant in science-society interactions. The following considerations thus focus on general issues of the science-society relation, acknowledging that the key questions would have to be adapted in a specific context.

2 Most sources used in this paper come from science and technology studies, philosophy of science and the transdisciplinarity/co-creation literature (see tab. 2). Note that a so-called "systematic" literature review may not be more authoritative, as the relevant terminologies are too heterogeneous to define sufficiently inclusive sets of keywords.

As a methodological note, it should be clear that this framework of dimensions and key questions is one of several ways to systematize the diversity of aspects that are relevant to the science-society relation (for alternative systematizations, see e.g. Lincoln & Guba, 2000; Fazey et al., 2018; Skelton, 2021). This follows straightforwardly from the ideal-typical approach discussed above. It should also be clear that systematizations of this sort cannot plausibly claim to be "authoritative" with respect to, e.g., the exact number and naming of dimensions, or the wording of the key questions. For instance, the fact that the following framework features six rather than, say, four or eight dimensions is a pragmatic compromise between complexity reduction and depth of detail. Furthermore, note that the differentiation between science-related and society-related dimensions does not mean that aspects that are part of the former are not social, or that aspects that are part of the latter have nothing to do with science; in fact, the question of whether these aspects can be separated neatly is answered quite differently by the SSIMs discussed in the next subsection. The differentiation should therefore be seen as one possible way of setting up these kinds of questions, not as a statement about the nature of science and society.

The science-related dimensions can be characterized as follows:

1. The dimension of epistemic standards refers to the norms, procedures and regulative ideals that constitute "good science". It also includes aspects of scientific quality assurance. Disagreements about the right epistemic standards constitute different views of how research processes should be designed. Key questions in this dimension refer to a range of procedural aspects of science, such as the ideal of value-freedom and its relation to objectivity, the integration of extra-scientific stakeholders into the research process, the role of extra-scientific (e.g. economic) influences on science, or quality assuring measures such as the peer review process.
2. The dimension of epistemic scope covers the epistemological and metaphysical principles of science: its limits, its reliability, its ability to find true and meaningful facts, and its relation to other types of knowledge. Disagreements in this dimension imply different ideas of what we can expect from science (science's "epistemic power"). Related key questions include issues of uncertainty and reliability, the nature of scientific and societal problems, the differences and similarities between scientific and non-scientific (local, traditional etc.) knowledge, or the question of whether scientific knowledge is socially contingent ("constructed").
3. The dimension of epistemic interests refers to the research problems that science should address, including the procedures by which research agendas should be determined. Disagreements in this dimension give rise to different views of how to dis-

cern the relevant research problems from the "vast oceans of truth that aren't worth exploring" (Kitcher, 2001, p. 148). Key aspects include questions such as whether research agendas should be determined autonomously by scientists, whether "pure" science has more or less value than applied science, whether basic and curiosity-driven research will yield practical applications in the long run, and how society should support the implementation of research agendas with financial and other resources.

The society-related dimensions can be characterized as follows:

4. The dimension of social legitimacy covers the procedural and substantial properties that constitute "good" decisions in societal contexts (e.g. in policy-making), including the authority that different actors should have in these contexts. Disagreements in this dimension imply diverging views of scientists' roles in public debates and decisions. The dimension includes the general question of what a practical decision needs to count as legitimate, but also more specific questions regarding the societal authority of scientists, the possibility and desirability of political neutrality, or the issue of whom or what scientists can represent in a decision-process.
5. The dimension of social rationalities addresses the nature of "good" practical reasons and their adoption by societal actors. The focus is on the conditions under which actors will act, or should act, on the basis of scientific results. As evidence-based action presupposes knowledge of the current state of science, this dimension also includes aspects of science communication. Disagreements in this dimension give rise to different views of rational decision-making and the public understanding of science. Crucial questions are how important evidence is for action (compared to, e.g., values), how research results should be communicated, and what role motivated reasoning plays for the acceptance of evidence by societal actors.
6. The dimension of social structure refers to the fundamental structure of modern societies, with science as one of many elements of that structure. Disagreements in this dimension lead to diverging ideas of how society is to be understood and how different societal spheres such as science, policy, or the general public can interact with each other. Key questions refer to the boundaries between science and society, the role of institutions (e.g. boundary organizations), the structural unity or fragmentation of society, the nature of conflict and power in society, or possibility of understanding and steering society from a global planner perspective.

4.2 A MODIFIED VERSION OF THE TDP TAXONOMY

The outlined dimensions describe a promising way to identify the fundamental or "philosophical" aspects of science-society interactions. As can be seen from the above list, however, these aspects are numerous, complex, and non-trivial. This is where taxonomic SSIMs come in: by suggesting ideal-typical answers to the above questions, they provide conceptual orientation and help to understand potential disagreements. As discussed before, these answers are not meant to be "true" or "false", but to serve as boundary cases that demarcate a conceptual space. This paper suggests a modified version of the TDP taxonomy for this purpose. As said before, the TDP taxonomy is not the only SSIM scheme on the market. However, while it would in principle be possible to take a different SSIM taxonomy as a starting point, the TDP taxonomy seems to be well suited due to its prominence and its conceptual breadth (particularly in comparison to dualist taxonomies such as the mode 1/2 distinction). Moreover, the TDP taxonomy provides a promising basis to include science-society aspects beyond Habermas' original focus on scientific policy advice. Besides this wider thematic focus, the following modification of the TDP taxonomy includes a more neutral terminology (without negative connotations) in order to ensure an open dialogue. As the relevant conceptual space may vary with the specifics of an interaction, these SSIMs may be further modified to address additional questions or philosophical positions. They may also be formulated in a less abstract and more context-specific way. Despite this flexibility, however, the following descriptions may serve as a solid starting point.

(T') The expert-centered SSIM modifies Habermas' technocratic model. In the dimension of epistemic standards, the SSIM holds that science is and should be value-free. This includes the research process itself, but also the determination of research agendas and the preparation of practical applications. Scientific quality is assured by peer review, which reliably filters out subjectivity and extra-scientific (e.g. financial) interests. Societal actors play no role in science. In the dimension of epistemic scope, the SSIM holds that science can solve most societal problems. Uncertainty is, if at all relevant, of transient nature. Science has privileged access to a context-independent truth, other kinds of knowledge are inferior. In the dimension of epistemic interests, the SSIM emphasizes basic research and scientific autonomy. Practical applications are thought to flow naturally from basic research. Society is to provide sufficient financial resources to science. In the dimension of social legitimacy, the SSIM is inspired by Plato's *Republic*, arguing that society is best governed by "lovers of that which is and of truth" (book VI, 501c). The SSIM

holds that that societal decisions are legitimate if they are objectively correct. As scientists possess knowledge about optimal solutions, they are legitimized to make societal decisions. Science represents both the facts and an unbiased ("objective") view of the common good in these decisions. In the dimension of social rationalities, the SSIM holds that action should be purely evidence-based. Societal actors are deemed ignorant and irrational. However, the SSIM also holds that society can be rationalized through top-down education. In the dimension of social structure, the SSIM pictures society as a complex machine that can be steered by a global planner. The boundaries between science and society are thought to be fixed and clear. Normative conflicts and power struggles are not expected.

(D') The decision-centered SSIM is adapted from Habermas' decisionist model. In the dimension of epistemic standards, the SSIM holds that value-freedom may not always be achieved, but that scientists can and should minimize value-ladenness effectively. Peer review helps to reduce values and extra-scientific (e.g. economic) influences. Societal stakeholders are not included into science. In the dimension of epistemic scope, the SSIM is confident that science can approximate truth, although some residual uncertainties might remain. While science is subject to historical change, it represents the best currently available knowledge. Consequently, science trumps other types of knowledge. However, science has nothing to say regarding normative issues. In the dimension of epistemic interests, the SSIM emphasizes society's freedom to grant or limit research funding without offering any justification. Society decides whether basic or applied research is more valuable. In the dimension of social legitimacy, the SSIM treats actors' autonomy is the highest good: societal actors are to make autonomous decisions, and science is to remain neutral regarding societal goals. In the dimension of social rationalities, the SSIM holds that actors' preferences are irrational, but that science can and should provide objective information about effective ways to realize these preferences. Values and facts are assumed to be cleanly distinguishable. However, societal actors are free to ignore scientific facts if they do not suit their values. In the dimension of social structure, the SSIM assumes fixed boundaries between science and society. Yet, society is not seen as a structural unity, but as a multiplicity of "various value spheres [that] stand in irreconcilable conflict with each other" (Weber, 1919/1958, p. 126). Science-society interactions are shaped by power and interests, and an integrated social planner perspective is viewed as impossible.

(P') The stakeholder-centered SSIM modifies Habermas' pragmatist model. In the dimension of epistemic standards, the SSIM holds that value-freedom is both impossible and undesirable. Societal actors are included in all stages of science. Qual-

ity assurance is realized through an extended peer review that involves societal stakeholders. In the dimension of epistemic scope, the SSIM argues that science is essentially uncertain, and that societal problems are unfit for engineering solutions ("wicked problems"). All knowledge is contingent and "constructed". Consequently, science is but one of many co-equal forms of knowledge. In the dimension of epistemic interests, the SSIM prioritizes applied research. Curiosity-driven research is strongly limited. Society decides about research funding, with a focus on useful and ethically welcomed research. In the dimension of social legitimacy, the SSIM holds that decisions are legitimate if they are ethical and result from a fair and inclusive deliberation. Scientists participate in these deliberations without any special authority. Political neutrality is assumed to be impossible. Rather, scientists take an advocatory role in societal debates (e.g. for sustainability). In the dimension of social rationalities, the SSIM holds that action-guiding values are inseparable from action-guiding facts. What counts as "rational" is highly contingent. Yet, actors can learn from each other, which also means that scientists and non-scientists should engage in mutual learning processes. In the dimension of social structure, the SSIM holds that society is an uncontrollable, highly heterogeneous multiplicity. Despite this diversity, conflicts can be resolved through respectful deliberation. Boundaries between science and society are assumed to be permeable and subject to change.

An overview of the proposed dimensions and ideal-typical SSIMs is given in table 1 and table 2.

Page 186:
Table 1: Generic dimensions and key questions of science-society interactions.

Page 187:
Table 2: Key assumptions of three ideal-typical taxonomic SSIMs (adapted from the TDP taxonomy).

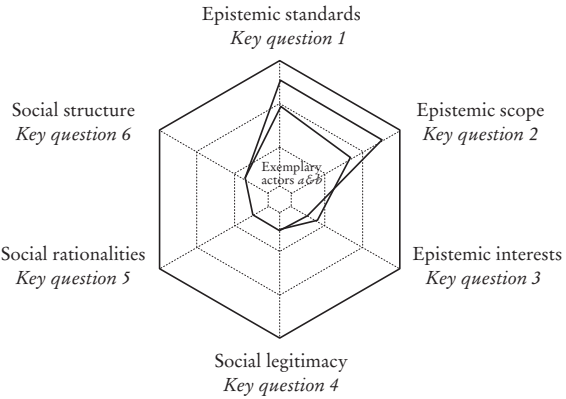
Dimensions	Subject	Key questions	Exemplary references	
Epistemic standards	The set of norms, procedures and regulative ideals that constitute 'good science'.	Can and should science be value-free? Should science involve societal stakeholders? How should scientific quality be assured? How should extra-scientific (e.g. economic) influences be addressed?	Merton (1942/1973), Lacey (1999), Longino (2002), Novotny (2003), Daston & Galison (2007), Carrier et al. (2008), Douglas (2009), Wilholt (2009), Kitcher (2011), Bremer & Meisch (2017).	Science-related dimensions
Epistemic scope	The epistemological and metaphysical principles of science.	How certain can scientific claims be? Can science solve social problems? Is science superior to other types of knowledge? Is science universally valid or historically and socially contingent?	Kuhn (1962), Latour & Woolgar (1979), Rittel & Webber (1973), Funtowicz & Ravetz (1993), Wynne (1996), Sarewitz (2004), Biddle (2013), Foyer & Kervan (2017).	
Epistemic interests	The problems, questions and challenges that science should address.	Should scientists enjoy freedom of research? What is the value of basic science, as compared to applied science? How should research resources be allocated? Will basic research lead to practical applications (in the long run)?	Bush (1945/1995), Polanyi (1962), Habermas (1971), Koertge (2000), Kitcher (2001), Kourany (2003), Wilholt (2010), Sarewitz (2016), Fazey et al. (2018).	
Social legitimacy	The procedural and substantial properties that constitute 'good' decisions.	What makes societal decisions legitimate? Whom/what can scientists represent in society? What role should scientists play in societal debates? Does epistemic authority imply practical authority?	Lackey (2007), Brown (2009), Pielke (2012), Turnhout et al. (2013), Wittmayer & Schäpke (2014), Edenhofer & Kowarsch (2015), Heazle & Kane (2016).	Society-related dimensions
Social rationalities	The nature of 'good' practical reasons and their adoption by societal actors.	Should practical decisions be evidence-based? What role do values play in decision-making? How should research results be communicated? What role does motivated reasoning play for the acceptance of evidence?	Douglas & Wildavsky (1983), Einsiedel (2000), Black (2001), Oreskes (2004), Sarewitz (2004), Sturgis & Allum (2004), Trench (2008), Kahan et al. (2010), Martin et al. (2020).	
Social structure	The fundamental structural principles of modern societies.	Are science-society boundaries fix or fluent? What role can/should boundary organizations play? How does power shape society-society relations? Can society be steered by a global planner?	Weber (1919/1958), Luhmann (1995), Weingart (1999), Guston (2001), Miller 2001, Dunlap & Brulle (2015), Grundmann & Rödder (2019).	

Dimensions	Expert-centered SSIM (T´)	Decision-centered SSIM (D´)	Stakeholder-centered SSIM (P´)	
Epistemic standards	Value-freedom is desirable and possible. No involvement of stakeholders. Quality assurance through standardized peer review.	Value-freedom is desirable, but not always possible. No involvement of stakeholders. Quality assurance through standardized peer review.	Value-freedom is undesirable and impossible. Societal stakeholders included in all stages of science. Quality assurance through extended peer review. Extra-scientific influences on science are inevitable.	Science-related dimensions
Epistemic scope	Uncertainty is irrelevant or transient. Science is able to solve most relevant problems. Science is superior to other forms of knowledge.	Uncertainty can be reduced, not completely eliminated. Science is powerful, but limited. Science only superior with regard to descriptive problems.	Science is essentially uncertain. Science cannot solve societal ('wicked') problems. Science is one of many co-equal forms of knowledge. Science is socially contingent ('constructed').	
Epistemic interests	Full freedom of research. Basic research is more valuable than applied research. Science chooses topics without accountability to society.	Limited freedom of research (societal funding needed). Society determines value of basic vs. applied science. Society allocates funds without accountability to science.	Low freedom of research (society co-determines topics). Applied research is more valuable than basic research. Science is accountable to society. Basic research is not needed for practical applications.	
Social legitimacy	Decisions are legitimate when objectively correct. Science represents facts and common good. Science takes decision-maker role.	Decisions are legitimate when agreed by societal actors. Science represents facts (but nothing more). Science takes advisory role.	Decisions legitimate when fair and ethically defensible. Science represents ethical concerns (e.g. sustainability). Science takes advocatory role. Epistemic authority does not imply practical authority.	Society-related dimensions
Social rationalities	Action should be purely evidence-based. Values are irrelevant for action. Societal actors are to be educated by science (top-down).	Action needs evidence and values (can be separated). Values are irrational. Societal actors are free to ignore scientific advice.	Action needs evidence and values (cannot be separated). Values (co-)determine goals and means of action. Bidirectional learning between science and society. Subjectivity is irreducible.	
Social structure	Science-society boundaries are clear and fixed. Society is a complex, but controllable machine. Science-society relation is shaped by reason, not power.	Science-society boundaries are clear and fixed. Society is a clash of subjective values and interests. Science-society relation is shaped by power.	Science-society boundaries are ambiguous and fluent. Society is a multiplicity of perspectives. Science-society relation is shaped by deliberation. Multiple societal spheres (no integrated perspective).	

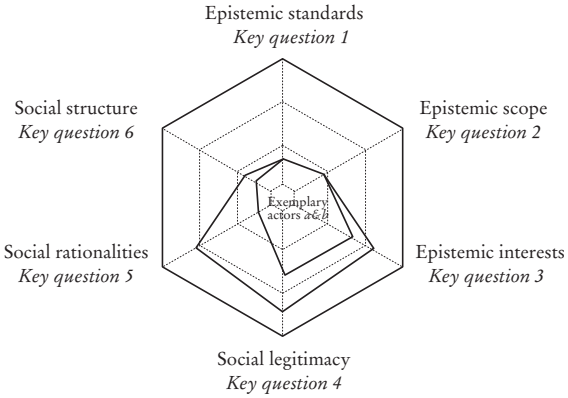
4.3 MAPPING ACTOR ASSUMPTIONS IN THE CONCEPTUAL SPACE: METHODOLOGICAL CONSIDERATIONS

In combination, the ideal-typical SSIMs and the thematic dimensions describe a conceptual space, i.e. a system of coordinates that helps to identify, compare and discuss actor assumptions. While the outlined tool has to be further developed and tested, the general idea should be clear from the previous sections: the tool would assess the attitudes that different actors have towards the key questions in each dimension; it would then determine the actors' proximity to the ideal-typical SSIMs, e.g. on a Likert scale (see figure 3 a-c); finally, the tool would determine convergences and disagreements between actors (see figure 3 d). This mapping of assumptions in the conceptual space could provide the ground for further discussions among actors and, ideally, help to design science-society interactions.

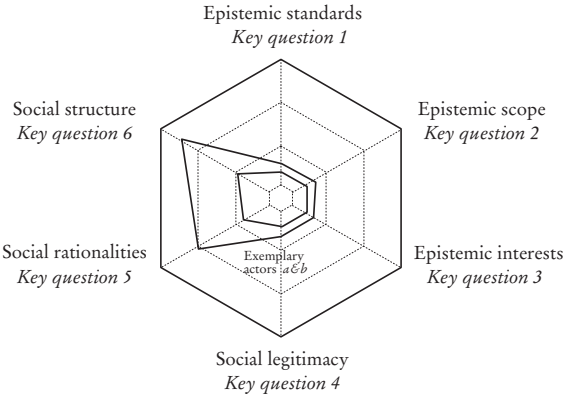
(3a) Expert-centered assumptions



(3b) Decision-centered assumptions



(3c) Stakeholder-centered assumptions



(3d) Conceptional space with actor assumptions

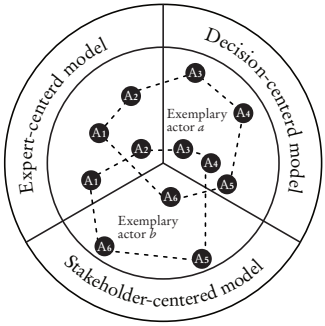


Figure 3: Mapping of actor assumptions in the conceptual space. Assumptions are identified by determining the actors' acceptance of each taxonomic SSIMs in each thematic dimension, e.g. on a Likert scale (a-c). Differences and convergences between actors are assessed by comparing the position of the actors' non-taxonomic SSIMs in the conceptual space (d).

The practical application of such a tool may be imagined in several ways. Note again that this paper is only a first step, and that the methodological specifics and the empirical testing must be elaborated in future work. Yet we can point towards some general considerations. For this, it is useful to distinguish between first-person, second-person, and third-person approaches (combinations are possible):

* In a first-person approach, actors may use the tool as a means of self-reflection, as well as a vehicle to make their assumptions explicit for an audience. Individual actors may find this helpful to get a clearer picture of their own background assumptions and to increase their reflexivity (Schwandt, 2011; Berger, 2015; Beck et al., 2021). The tool may also be relevant for collective actors (institutions, organizations etc.) who wish to inform their stakeholders (clients, partners, or the public) about their science-society assumptions. We could even imagine a standardization of such a self-assessment tool in a given context. If a number of institutions would agree on a specific formulation of the tool in their concrete context, and if each of these institutions would perform the self-assessment and publish the results, stakeholders could compare the SSIMs used by these institutions in a transparent manner. This could be a

valuable service to the stakeholders, and may spark a fruitful debate in the respective context.

* In a second-person approach, the tool may be used as a means of group reflection, e.g. in a workshop. Such an approach may be sensible in contexts where scientists and non-scientists collaborate over long periods of time (transdisciplinary projects, mixed committees etc.). Contrary to a first-person approach, where actors merely assess and communicate their assumptions, this approach is more deliberative. Actors would start by collectively interpreting and modifying the key questions and ideal-typical SSIMs in their context. They would then locate themselves in the resulting conceptual space. The main part would consist in a deliberation about controversial assumptions. After that, actors may use the tool again to identify possible changes in beliefs. Finally, actors would discuss the practical implications for their collaborative projects. An important source for developing the tool into this direction is the *Toolbox Dialogue* approach (Eigenbrode et al., 2007; Hubbs et al., 2021; Laursen et al., 2021), a workshop-based method in which "members of [cross-disciplinary] teams explore the implicit beliefs and values that influence their project contributions" (Hubbs et al., 2021, p. xiii). Evaluations of Toolbox workshops are promising (Rinkus and O'Rourke, 2021; Robinson and Gonnerman, 2021), which justifies hopes that the tool envisioned in this paper may be applicable in a second-person approach. Similarly, there is a literature on deliberative citizen panels in the science-society field that offers further methodological inspiration (e.g. Bertrand et al., 2017).
* In a third-person approach, the tool may be used to identify actors' assumptions from an outside perspective. Rather than initiating a deliberation or assessing one's own assumptions, this type of approach focusses on gathering and interpreting data. This may be useful for empirical researchers, but also for institutions or project organizers who seek information about their stakeholders' expectations. These data would reveal how broad the spectrum of assumptions held in a given target group actually is, and may help institutions or organizers to identify trade-offs and synergies in their stakeholder activities. The fact that similar data have been gathered in other studies (Steel et al., 2004; Gray and Campbell, 2008; Reiners et al., 2013; Steel et al., 2017; Van der Hel, 2018) shows that actors are prepared to answer the tool's key questions (in questionnaires, interviews etc.) despite their abstract and philosophical character. Additionally, the tool may be used to generate empirical hypotheses about actor assumptions. These hypotheses could be tested either directly, i.e. by surveying the respective actors, or indirectly through observation, discourse analysis, or other methods.

5. Clonclusion and open questions

This paper has discussed the strengths and weaknesses of SSIMs, focusing on the popular distinction between technocratic, decisionist and pragmatist models (the TDP taxonomy). It has argued that SSIMs and SSIM taxonomies are valuable reducers of complexity and illuminators of conceptual divides. However, these merits are undermined by a tendency in the science-society literature to treat taxonomic SSIMs as antagonistic theoretical camps or as representations of actor beliefs. To avoid the "straw man" debates that result from this tendency, the paper has argued that we should put more emphasis on the ideal-typical and heuristic nature of SSIMs. Starting from this interpretation, the paper has presented some tentative ideas for a reflexive tool. This tool consists of six dimensions of key questions and a modified version of the TDP taxonomy. It uses the TDP models as boundary cases that demarcate a conceptual space. By locating themselves in this conceptual space, actors (scientists and non-scientists) can identify, compare and discuss their implicit assumptions. The results may be used to improve science-society interactions in diverse contexts, e.g. in advisory processes or stakeholder encounters within research projects.

It may be objected that that the envisioned tool is too abstract to address real-world science-society interactions. However, this paper has not claimed that the tool can be readily applied. It should rather be seen as a starting point from which more context-specific versions can be developed. Another objection may be that actor assumptions can be identified without taxonomic SSIMs, just using the key questions. Note, however, that this paper has not described taxonomic SSIMs as *necessary*, but as *beneficial*. Actors may find it easier to unwrap their own assumptions if they can indicate the degree to which they accept or reject a clear-cut narrative. By locating their assumptions in a common conceptual space, taxonomic SSIMs may also help actors understand the extent and gravity of potential disagreements. A third, connected objection is that the tool does not show how such disagreements can be resolved (see e.g. Failing et al., 2007; Laursen et al., 2021). This point is well taken. However, transparency about fundamental assumptions is a promising starting ground, and may even be regarded as a value in itself (Elliott & McKaughan, 2013; Elliott & Resnik, 2014). Thus, while the issue of potential disagreements is indeed crucial, it does not speak against the tool.

Regarding the theoretical part of this paper, there are two obvious, yet opposing objections: some may argue that taxonomic SSIMs are more than ideal-types, namely actual

bearers of truth and falsity; others may argue that, quite on the contrary, taxonomic SSIMs have always been regarded as ideal-types, which makes the ideal-typical interpretation somewhat trivial. While this paper disagrees with both objections, they point towards a crucial ambivalence in the science-society literature. It is true that taxonomic SSIMs have often been described as ideal-types (Heinrichs, 2005; Lompe, 2006; Wittmayer & Schäpke, 2014; Kowarsch, 2016; Jahn et al., 2021). As argued before, however, this point has not been systematically elaborated in the science-society literature so far. Also, the point is counteracted by the discussed tendencies in the science-society literature, such as the popular idea that certain actors are "clearly" (Beck 2011, p. 298) guilty of using a false taxonomic SSIM. The insight that SSIMs should not be seen as antagonistic theory camps is also counteracted by polemics against certain taxonomic SSIMs, as well as the widespread discomfort that these taxonomic SSIMs are "refusing to die despite so many mortal blows" (Durant, 2016, p. 31, see also Pielke, 2012, p. 8; Van der Hel, 2018, p. 256). Finally, if the ideal-typical interpretation were trivial, worries that taxonomic SSIMs are "straw-men" would be unnecessary. Still several authors voiced exactly this worry (Sturgis & Allum, 2004; Trench, 2008; Martin, 2012).

This motivates the conjecture that, rather than being regarded as ideal-types, taxonomic SSIMs are in fact often seen as antagonistic theory camps and as representations of actor beliefs in the science-society literature. This paper has argued that this is an unproductive perspective. Note, however, that this paper has not claimed that all SSIMs are equally convincing, philosophically speaking, but that this may not be a *fruitful question* – at least not if the aim is to provide conceptual orientation and to enable an open discussion among the participants of a science-society interaction. This paper should therefore not be taken to defend some sort of SSIM relativism; rather, the idea is that taxonomic SSIMs, together with the six dimensions of key questions, provide us with a way in which actor assumptions can be identified and then, in a subsequent debate, be constructively discussed among actors. Furthermore, the paper has not argued that it is strictly *impossible* for an actor to subscribe to all assumptions of a taxonomic SSIM in their purest form. The point is rather that this is not typical, and that we should not focus on such extreme cases. As a final remark, note that the theoretical and practical claims of this paper are to some degree independent: even if one insists that a taxonomic SSIM is plain false, one can still use this SSIM as a heuristics to identify actor assumptions. Vice versa, even if one believes that the envisioned tool is unfit for real-world application, the theoretical considerations may still enrich one's understanding of SSIMs and the science-society relation.

Acknowledgements

The greatest part of this work was conducted and financed within the framework of the Helmholtz Institute for Climate Service Science (HICSS), a cooperation between Climate Service Center Germany (GERICS) and Universität Hamburg, Germany. Additional funding was provided by the Research Unit Sustainability and Climate Risks at Universität Hamburg, Germany. I would like to thank Hermann Held, Susanne Schuck-Zöller, Mirko Suhari, Laura Schmidt, Torsten Wilholt, Dietmar Hübner and Mareike Blum for valuable feedback.

References

Bacon, F. (2000). *Francis Bacon: The New Organon*. Cambridge University Press. (First publ. 1620)

Beck, S. (2011). Moving beyond the linear model of expertise? IPCC and the test of adaptation. *Regional Environmental Change*, 11(2), 297–306. https://doi.org/10.1007/s10113-010-0136-2

Beck, J. M., Elliott, K. C., Booher, C. R., Renn, K. A., & Montgomery, R. A. (2021). The application of reflexivity for conservation science. *Biological Conservation*, 262, 109322. https://doi.org/10.1016/ j.biocon.2021.109322

Berger, R. (2015). Now I see it, now I don't: researcher's position and reflexivity in qualitative research. *Qualitative Research*, 15(2), 219–234. https://doi.org/10.1177/1468794112468475

Bertrand, P., Pirtle, Z., & Tomblin, D. (2017). Participatory technology assessment for Mars mission planning: Public values and rationales. *Space Policy*, 42, 41–53. https://doi.org/10.1016/ j.spacepol.2017. 08.004

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, 3(2), 207–220. https://doi.org/10.1007/s13194-012-0062-x

Biddle, J. (2013). State of the field: Transient underdetermination and values in science. *Studies in History and Philosophy of Science*, 44(1), 124–133. https://doi.org/10.1016/j.shpsa.2012.09.003

Black, N. (2001). Evidence based policy: proceed with care. *BMJ*, 323(7307), 275–279. https://doi.org/ 10.1136/bmj.323.7307.275

Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In R.L. Launer, & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). New York: Academic Press.

Bremer, S., & Meisch, S. (2017). Co-production in climate change research: reviewing different perspectives. *WIREs Climate Change*, 8(6). https://doi.org/10.1002/wcc.482

Brossard, D., & Lewenstein, B. V. (2010). A Critical Appraisal of Models of Public Understanding of Science Using Practice to Inform Theory. In L. Kahlor, & P. Stout (Eds.), *Communicating Science: New Agendas in Communication* (pp. 11–39). New York.

Brown, M. B., Lentsch, J., & Weingart, P. (2005). Representation, Expertise, and the German Parliament: A Comparison of Three Advisory Institutions. In S. Maasen, & P. Weingart (Eds.), *Democratization of Expertise. Exploring Novel Forms of Scientific Advice in Political Decision-Making* (pp. 81–100). Springer.

Brown, M. B. (2009). *Science in democracy. Expertise, institutions, and representation*. MIT Press.

Bush, V. (1995). *Science, The Endless Frontier* (Reprint). Ayer Company Publishers. (First publ. 1945)

Carrier, M., Howard, D., & Kourany, J. A. (2008). The challenge of the social and the pressure of practice. Science and values revisited. University of Pittsburgh Press.

Colyvan, M. (2013). Idealisations in normative models. Synthese, 190(8), 1337–1350. https:// doi.org/ 10.1007/s11229-012-0166-z

Crowley, S. J., & O'Rourke, M. (2021). Communication Failure and Cross- Disciplinary Research. In G. Hubbs, M. O'Rourke, & S.H. Orzack (Eds.), *The toolbox dialogue initiative: The power of cross-disciplinary practice* (pp. 1–16). CRC Press.

Daston, L., & Galison, P. (2007). *Objectivity*. Zone Books.

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Douglas, M., & Wildavsky, A. (1983). Risk and culture. An essay on the selection of technological and environmental dangers. University of California Press.

Dunlap, R. E., & Brulle, R. J. (Eds.) (2015). *Climate change and society. Sociological perspectives*. Oxford University Press.

Durant, D. (2016). The undead linear model of expertise. In M. Heazle, & J. Kane (Eds.), *Policy Legitimacy, Science and Political Authority. Knowledge and action in liberal democracies* (pp. 17–37). Routledge.

Edenhofer, O., & Kowarsch, M. (2015). Cartography of pathways: A new model for environmental policy assessments. *Environmental Science & Policy*, 51, 56–64. https://doi.org/10.1016/j.envsci.2015.03.017

Edenhofer, O., & Seyboth, K. (2013). Intergovernmental Panel on Climate Change (IPCC). In J. Shogren (Eds.), *Encyclopedia of Energy, Natural Resource, and Environmental Economics* (pp. 48–56). Elsevier. https://doi.org/10.1016/B978-0-12-375067-9.00128-5

Eigenbrode, S. D., O'Rourke, M., Wulfhorst, J. D., Althoff, D. M., Goldberg, C. S., Merrill, K., Morse, W., Nielsen-Pincus, M., Stephens, J., Winowiecki, L., & Bosque-Pérez, N.A. (2007). Employing Philosophical Dialogue in Collaborative Science. *BioScience*, 57(1), 55–64. https://doi.org/10.1641/ B570109

Einsiedel, E. (2000). Understanding 'publics' in the public understanding of science. In M. Dierkes, & C. von Grote (Eds.), *Between understanding and trust: The public, science and technology* (pp. 205–216). OPA.

Elliott, K. C. (2011). *Is a little pollution good for you? Incorporating societal values in environmental research*. Oxford: Oxford University Press.

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic Values and the Multiple Goals of Science. *Philosophy of Science*, 81(1), 1–21. https://doi.org/10.1086/674345

Elliott, K. C., & Resnik, D. B. (2014). Science, policy, and the transparency of values. *Environmental Health Perspectives*, 122(7), 647–650. https://doi.org/10.1289/ehp.1408107

Estlund, D. (2003). Why Not Epistocracy? In N. Reshotko (Ed.), *Desire, Identity, and Existence. Essays in Honor of T.M. Penner* (pp. 53–69). Academic Printing & Publishing.

Failing, L., Gregory, R., & Harstone, M. (2007). Integrating science and local knowledge in environmental risk management: A decision-focused approach. *Ecological Economics*, 64(1), 47–60. https://doi.org/ 10.1016/j.ecolecon.2007.03.010

Fazey, I., Schäpke, N., Caniglia, G., Patterson, J., Hultman, J., van Mierlo, B., Säwe, F., Wiek, A., Wittmayer, J., Aldunce, P., Al Waer, H., Battacharya, N., Bradbury, H., Carmen, E., Colvin, J., Cvitanovic, C., D'Souza, M., Gopel, M., Goldstein, B., ..., & Wyborn, C. (2018). Ten essentials for action-oriented and second order energy transitions, transformations and climate change research. *Energy Research & Social Science*, 40, 54–70. https://doi.org/10.1016/j.erss.2017.11.026

Felt, U., Wynne, B., Callon, M., Gonçalves, M. E., Jasanoff, S., Jepsen, M., Joly, P.B., Konopasek, Z., May, S., Neubauer, C., Rip, A., Siune, K., Stirling, A., & Tallacchini, M. (2007). *Taking European knowledge society seriously. Report of the expert group on science and governance to the Science, Economy and Society Directorate, Directorate-General for Research, European Commission*. Office for Official Publications of the European Communities. https://op.europa.eu/en/publication-detail/-/publication/5d0e77c7-2948-4ef5-aec7-bd18efe3c442

Foyer, J., & Kervran, D. D. (2017). Objectifying traditional knowledge, re-enchanting the struggle against climate change. In S. C. Aykut (Ed.), *Globalising the climate: COP21 and the climatisation of global debates* (pp. 153–172). Routledge.

Frigg, R., & Hartmann, S. (2020). Models in Science. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). https://plato.stanford.edu/archives/spr2020/entries/models-science/

Funtowicz, S. O., & Ravetz, J. R. (1993). Science for the post-normal age. *Futures*, 25(7), 739–755. https://doi.org/10.1016/0016-3287(93)90022-L

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge. The Dynamics of Science and Research in Contemporary Societies*. Sage.

Giere, R. N. (2004). How Models Are Used to Represent Reality. *Philosophy of Science*, 71(5), 742–752. https://doi.org/10.1086/425063

Gluckman, P. D. (2011). *Towards better use of evidence in policy formation. A discussion paper*. Office of the Prime Minister's Science Advisory Committee. https://dpmc.govt.nz/ sites/default/files/2021-10/pmcsa-Towards-better-use-of-evidence-in-policy-formation.pdf

Gray, N. J., & Campbell, L. M. (2009). Science, policy advocacy, and marine protected areas. *Conservation Biology*, 23(2), 460–468. https://doi.org/10.1111/j.1523-1739.2008.01093.x

Grundmann, R., & Rödder, S. (2019). Sociological Perspectives on Earth System Modeling. *Journal of Advances in Modeling Earth Systems*, 11(12), 3878–3892. https://doi.org/10.1029/ 2019MS001687

Grunwald, A., & Saretzki, T. (2020). Demokratie und Technikfolgenabschätzung. *TATuP*, 29(3), 11–17. https://doi.org/10.14512/tatup.29.3.11

Guston, D. (2001). Boundary organizations in environmental policy and science: An introduction. *Science, Technology, & Human Values*, 26, 399–408. https://doi.org/10.1177/ 016224390102600401

Habermas, J. (1970). *Toward a rational society. Student protest, science, and politics*. Beacon Press.

Habermas, J. (1971). *Knowledge and human interests*. Beacon Press.

Hagedorn, G., Loew, T., Seneviratne, S. I., Lucht, W., Beck, M. L., Hesse, J., Knutti, R., Quaschning, V., Schleimer, J. H., Mattauch, L., Breyer, C., Hübener, H., Kirchengast, G., Chodura, A., Clausen, J., Creutzig, F., Darbi, M., Daub, C. H., Ekardt, F., … , & Zens, J. (2019). The concerns of the young protesters are justified: A statement by Scientists for Future concerning the protests for more climate protection. *GAIA*, 28(2), 79–87. https://doi.org/10.14512/gaia.28.2.3

Harding, S. (1995). "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3), 331–349. https://doi.org/10.1007/BF01064504

Heazle, M., & Kane, J. (Eds.). (2016). *Policy legitimacy, science and political authority. Knowledge and action in liberal democracies*. Routledge.

Heazle, M., Kane, J., & Patapan, H. (2016). Good public policy. On the interaction of political and expert authority. In M. Heazle, & J. Kane (Eds.), *Policy Legitimacy, Science and Political Authority Knowledge and action in liberal democracies* (pp. 1–16). Routledge.

Heinrichs, H. (2005). Advisory Systems in Pluralistic Societies: A Criteria-Based Typology to Assess and Optimize Environmental Policy Advice. In S. Maasen, & P. Weingart (Eds.), *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making* (pp. 41–61). Springer.

Hessels, L. K., van Lente, H., & Smits, R. (2009). In search of relevance: the changing contract between science and society. *Science and Public Policy*, 36(5), 387–401. https://doi.org/10.3152/ 030234209X442034

Hoyningen-Huene, P. (2009). Tensions Between Science And Society. *Axiomathes*, 19(4), 417–424. https://doi.org/10.1007/s10516-009-9088-x

Hubbs, G., O'Rourke, M., & Orzack, S. H. (Eds.). (2021). *The toolbox dialogue initiative. The power of cross-disciplinary practice*. CRC Press.

Hulme, M. (2009). *Why we disagree about climate change. Understanding controversy, inaction and opportunity*. Cambridge University Press.

Jahn, S., Newig, J., Lang, D. J., Kahle, J., & Bergmann, M. (2022). Demarcating transdisciplinary research in sustainability science—Five clusters of research modes based on evidence from 59 research projects. *Sustainable Development*, 30(2), 343–357. https://doi.org/10.1002/sd.2278

Jasanoff, S. (1990). *The fifth branch. Science advisers as policymakers*. Harvard University Press.

Jasanoff, S. (2003). (No?) Accounting for expertise. *Science and Public Policy*, 30(3), 157–162. https://doi.org/10.3152/147154303781780542

John, S. (2015). Inductive risk and the contexts of communication. *Synthese*, 192(1), 79–96. https://doi.org/10.1007/s11229-014-0554-7

Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, 14(2), 147–174. https://doi.org/10.1080/13669877.2010.511246

Kitcher, P. (2001). *Science, truth, and democracy*. Oxford University Press.

Kitcher, P. (2011). *Science in a democratic society*. Prometheus Books.

Koertge, N. (2000). Science, Values, and the Value of Science. *Philosophy of Science*, 67(S3), S45–S57. https://doi.org/10.1086/392808

Kourany, J. A. (2003). A Philosophy of Science for the Twenty-First Century. *Philosophy of Science*, 70(1), 1–14. https://doi.org/10.1086/367864

Kowarsch, M. (2016). *A Pragmatist Orientation for the Social Sciences in Climate Policy. How to Make Integrated Economic Assessments Serve Society*. Springer International Publishing.

Kowarsch, M., Garard, J., Riousset, P., Lenzi, D., Dorsch, M., Knopf, B., Harrs, J., & Edenhofer, O. (2016). Scientific assessments to facilitate deliberative policy learning. *Palgrave Communications*, 2(1), 1–20. https://doi.org/10.1057/palcomms.2016.92

Krishna, V. V. (2014). Changing Social Relations between Science and Society: Contemporary Challenges. *Science, Technology and Society*, 19(2), 133–159. https://doi.org/10.1177/ 0971721814529876

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lacey, H. (1999). *Is science value free? Values and scientific understanding*. Routledge.

Lackey, R. T. (2007). Science, scientists, and policy advocacy. *Conservation Biology*, 21(1), 12–17. https://doi.org/10.1111/j.1523-1739.2006.00639.x

Latour, B., & Woolgar, S. (1979). *Laboratory life. The social construction of scientific facts*. Sage.

Laursen, B. K., Gonnerman, C., & Crowley, S. J. (2021). Improving philosophical dialogue interventions to better resolve problematic value pluralism in collaborative environmental science. *Studies in History and Philosophy of Science*, 87, 54–71. https://doi.org/10.1016/j.shpsa.2021.02.004

Lincoln, Y. S., & Guba, E. G. (2000). Paradigmatic Controversies, Contradictions, and Emerging Controversies. In N. Denzin, & Y. S. Lincoln (Eds.), *The Sage Handbook of Qualitative Research* (pp. 163–188). Sage.

Lompe, K. (2006). Traditionelle Modelle der Politikberatung. In S. Falk, D. Rehfeld, A. Römmele, & M. Thunert (Eds.), *Handbuch Politikberatung*. VS Verlag für Sozialwissenschaften. https://doi.org/ 10.1007/978-3-531-90052-0_3

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Luhmann, N. (1995). *Social systems*. Stanford University Press.

Martin, B. R. (2012). Are universities and university research under threat? Towards an evolutionary model of university speciation. *Cambridge journal of economics*, 36(3), 543–565. https://doi.org/ 10.1093/cje/bes006

Martin, G. P., Hanna, E., McCartney, M., & Dingwall, R. (2020). Science, society, and policy in the face of uncertainty: reflections on the debate around face coverings for the public during COVID-19. *Critical Public Health*, 30(5), 501–508. https://doi.org/10.1080/09581596.2020.1797997

Merton, R. K. (1973). The normative structure of science. In *The sociology of science. Theoretical and empirical investigations* (pp. 267–278). University of Chicago Press. (First publ. 1942)

Miller, C. (2001). Hybrid Management: Boundary Organizations, Science Policy, and Environmental Governance in the Climate Regime. *Science, Technology, & Human Values*, 26(4), 478–500. https://doi.org/10.1177/016224390102600405

Millstone, E. (2005) Analysing the role of science in public policy-making. In P. van Zwanenberg, & E. Millstone, *BSE: Risk, Science and Governance* (pp. 11–38). Oxford University Press.

Morgan, M. S., & Morrison, M. (Eds.). (1999). *Models as mediators. Perspectives on natural and social sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9780511660108

Nielsen, L. A. (2001). Science and Advocacy Are Different – And We Need to Keep Them That Way. *Human Dimensions of Wildlife*, 6(1), 39–47. https://doi.org/10.1080/10871200152668689

Nowotny, H. (2003). Democratising expertise and socially robust knowledge. *Science and Public Policy*, 30(3), 151–156. https://doi.org/10.3152/147154303781780461

Nowotny, H, Scott, P., & Gibbons, M. (2003). Introduction. 'Mode 2' revisited: The new production of knowledge. *Minerva*, 41(3), 179–194.

Oreskes, N. (2004). Science and public policy: what's proof got to do with it? *Environmental Science & Policy*, 7(5), 369–383. https://doi.org/10.1016/j.envsci.2004.06.002

Pielke, R. A. (2012). *The honest broker. Making sense of science in policy and politics* (8th ed.). Cambridge University Press.

Plato (1991). *The republic of Plato* (2nd ed., transl. and ed. by A. Bloom). Basic Books. (First publ. ca. 375 BC)

Polanyi, M. (1962). The Republic of science. *Minerva*, 1(1), 54–73. https://doi.org/10.1007/ BF01101453

Reiners, D. S., Reiners, W. A., & Lockwood, J. A. (2013). The relationship between environmental advocacy, values, and science: a survey of ecological scientists' attitudes. *Ecological Applications*, 23(5), 1226–1242. https://doi.org/10.1890/12-1695.1

Rinkus, M. A., & O'Rourke, M. (2021): Qualitative Analyses of the Effectiveness of Toolbox Dialogues. In G. Hubbs, M. O'Rourke, & S. H. Orzack (Eds.), *The toolbox dialogue initiative: The power of cross-disciplinary practice* (pp. 142–161). CRC Press.

Rittel, H. W. J., & Webber, M. M. (1973): Dilemmas in a general theory of planning. *Policy Science*, 4, 155–169. https://doi.org/10.1007/BF01405730

Robinson, B., & Gonnerman, C. (2021): Enhancing Cross-Disciplinary Science through Philosophical Dialogue Evidence of Improved Group Metacognition for Effective Collaboration. In G. Hubbs, M. O'Rourke, & S. H. Orzack (Eds.), *The toolbox dialogue initiative: The power of cross-disciplinary practice* (pp. 127–141). CRC Press.

Sarewitz, D. (2004). How science makes environmental controversies worse. *Environmental Science & Policy*, 7(5), 385–403. https://doi.org/10.1016/j.envsci.2004.06.001

Sarewitz, D. (2016) Saving Science. *The New Atlantis*, 49, 4–40.

Schmidt, J. C. (2011). What is a problem? On problem-oriented interdisciplinarity. *Poiesis & Praxis*, 7(4), 249–274. https://doi.org/10.1007/s10202-011-0091-0

Schwandt, T. A. (2011) Reflexivity. In *The SAGE Dictionary of Qualitative Inquiry*. Sage. https://dx.doi.org/10.4135/9781412986281.n298

Skelton, M. (2021). Orders of Social Science: Understanding Social-Scientific Controversies and Confluence on What "High-Quality" Knowledge and "Good" Adaptation Is. *Frontiers in Climate*, 3, 589265. https://doi.org/10.3389/fclim.2021.589265

Steel, B., List, P., Lach, D., & Shindler, B. (2004). The role of scientists in the environmental policy process: a case study from the American west. *Environmental Science & Policy*, 7(1), 1–13. https://doi.org/10.1016/j.envsci.2003.10.004

Steel, D., Gonnerman, C., & O'Rourke, M. (2017). Scientists' attitudes on science and values: Case studies and survey methods in philosophy of science. *Studies in History and Philosophy of Science*, 63, 22–30. https://doi.org/10.1016/j.shpsa.2017.04.002

Sturgis, P., & Allum, N. (2004). Science in Society: Re-Evaluating the Deficit Model of Public Attitudes. *Public Understanding of Science*, 13(1), 55–74. https://doi.org/10.1177/0963662504042690

Trench, B. (2008). Towards an Analytical Framework of Science Communication Models. In D. Cheng, M. Claessens, T. Gascoigne, J. Metcalfe, B. Schiele, & S. Shi (Eds.), *Communicating Science in Social Contexts* (pp. 119–135). Springer. https://doi.org/10.1007/978-1-4020-8598-7_7

Turnbull, D. (1993). *Maps are territories. Science is an atlas*. University of Chicago Press.

Turnhout, E., Stuiver, M., Klostermann, J., Harms, B., & Leeuwis, C. (2013). New roles of science in society: Different repertoires of knowledge brokering. *Science and Public Policy*, 40(3), 354–365. https://doi.org/10.1093/scipol/scs114

Van der Hel, S. (2018). Science for change: A survey on the normative and political dimensions of global sustainability research. *Global Environmental Change*, 52, 248–258. https://doi.org/10.1016/ j.gloenvcha.2018.07.005

Weber, M. (1949). "Objectivity" in Social Science and Social Policy. In *On the Methodology of the Social Sciences* (pp. 50–112). The Free Press. (First publ. 1904)

Weber, M. (1958). Science as a Vocation. *Daedalus*, 87(1), 111–134. (First publ. 1919)

Weingart, P. (1999). Scientific expertise and political accountability: paradoxes of science in politics. *Science and Public Policy*, 26(3), 151–161. https://doi.org/10.3152/147154399781782437

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science*, 40(1), 92–101. https://doi.org/10.1016/j.shpsa.2008.12.005

Wilholt, T. (2010). Scientific freedom: its grounds and their limitations. *Studies in History and Philosophy of Science*, 41(2), 174–181. https://doi.org/10.1016/j.shpsa.2010.03.003

Wittmayer, J. M., & Schäpke, N. (2014). Action, research and participation: roles of researchers in sustainability transitions. *Sustainability Science*, 9(4), 483–496. https://doi.org/10.1007/s11625-014-0258-4

Wynne, B. (1996). May the Sheep Safely Graze? A Reflexive View of the Expert-Lay Knowledge Divide. In S. M. Lash, B. Szerszynski, & B. Wynne (Eds.), *Risk, Environment and Modernity. Towards a New Ecology* (pp. 44–83). Sage.