Descriptive Practices and Values in Endocrine Disruption Research

A Dissertation

SUBMITTED TO THE FACULTY OF

UNIVERSITY OF MINNESOTA

BY

John Douglas Powers

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Alan Love

August 2016

Acknowledgements

I wish to thank my parents, Martha and Billy Carr, and Doug and Geni Powers, for their love and support. Thanks to my advisor, Alan Love, for being a great coach. Thanks to my committee, Naomi Scheman, Ruth Shaw, Ken Waters, and Bill Wimsatt for their encouragement and criticisms. Thanks to Sandra Peterson and Geoffrey Hellman for comments on Chapter 2. Thanks to Heather Douglas, Ingo Brigandt, Melanie Bowman, Esther Rosario, Kevin Elliott, and Ted Richards for comments on Chapter 3. Thanks to the benefactors of the Tom Lapic Memorial Fellowship, the Tan Spark Fellowship, and the Douglas E. Lewis Fellowship for supporting my research. Thanks to the Navarro lab at Universidad Nacional Autónoma de México for coffee and a workspace. Thanks to Mary and Nathan Gass for the laughs, love, and airport shuttling. Thanks to Will Bausman and Matthew Ruble for good conversations and great climbs. Thanks to my grandparents, Pat and Snook Powers, and Evelyn Parker for their generosity and love. Thanks to the Chávez and Huelgas families, and Doña Marie, for making me feel at home. Thanks to the BIG and BIG-WIG discussion groups, and the Minnesota Center for Philosophy of Science for an education that cannot be found anywhere else. Thanks to Anita Wallace, Judy Grandbois, and Pam Groscost for keeping me on track. Thanks to Valerie Tiberius for her leadership and kind words. Thanks to Michael Calasso for the pasta and commiseration. Thanks to the Clarkrange cabin crew for the tough love and for always welcoming me home. Thanks to Patty Derycke for her love and wisdom, and to Bob Derycke for introducing me to the joy of dialogue. Thanks to John Nolt and Lee Shepski for helping me to pursue philosophy professionally. Lee, you are missed.

Finally, thanks to Gabriela Huelgas Morales, for her sweetness, brilliance, and grace (and the figures in Chapter 4). *Te necesito todos los días*.

Dedication

This thesis is dedicated to scientists who work for the sake of human wellbeing and protection of the environment.

Abstract

This work is a philosophical analysis of descriptive practices and values in endocrine disruption research. Chapter 1 provides an accessible overview. In Chapter 2, I develop a nonreductionist epistemology of research into the endocrine disrupting properties of the herbicide atrazine. I argue that criteria of adequacy governing descriptive practices in atrazine research serve to help organize and coordinate the activities and contributions of researchers from diverse disciplinary backgrounds. In Chapter 3, I examine the influence of non-epistemic values on terminology choice in endocrine disruption research. Researchers face choices about whether or not to use gendered language to describe the harmful effects of atrazine. I argue that such choices are locations of "inductive risk." In Chapter 4, I examine traditional "global demarcation" approaches for recognizing science that is problematically value-laden. I argue that global demarcation projects as currently undertaken are unlikely to meet their aims and suggest an alternative approach. This alternative approach reinterprets global demarcation projects as providing prima facie principles of good science. The prima facie principles resulting from such modest demarcation projects are to be integrated with appeals to local criteria of adequacy for scientific practices, and principles of inference for illicit influences of values in science. I illustrate this approach using a case of industry funded pesticide research. In Chapter 5, I argue that choices about whether to be a monist or pluralist about scientific terms depend on the epistemic and nonepistemic goals and values of debate participants. I illustrate by analyzing monism and pluralism about the terms 'potency' and 'endocrine disruptor' in recent endocrine disruption debates.

Table of Contents

List of Tables

List of Figures

Chapter 1: Overview

This dissertation is comprised of four relatively independent essays analyzing descriptive practices and values in endocrine disruption research. Endocrines, or hormones, are molecules that act as messengers between the various tissues and organs of biological organisms. Although scientists have recognized since the early 20th century that substances from outside the body can impact the function of endocrine systems, modern endocrine disruption research began in the early 1990s with the work of Theo Colborn and her colleagues. Colborn was puzzled by a wide range of reproductive, behavioral, and developmental abnormalities in wildlife that were cropping up regularly in the scientific literature and in the reports of amateur naturalists. These abnormalities did not seem to be explained by the presence of any then-recognized toxins. Colburn hypothesized that manmade chemicals that were not then known to be toxic might be interfering with the endocrine systems of wildlife and thereby causing the abnormalities. Colburn thought that these same chemicals were also likely to be acting on humans. She claimed that traditional toxicological tests were missing a large class of chemicals that posed serious health risks to humans and the environment.

The publication of Colburn's early work has sparked 25 years of intense research and controversy. Although governments and NGOs have implemented endocrine disruptor screening and testing programs, and issued reports on the impacts of endocrine

disruptors, the goals and standards of these programs and the findings of these reports have been contested at every turn. Traditional toxicologists have been resistant to making the methodological revisions recommended by Colburn and her allies. Emerging endocrinological approaches to toxicology have suggested that even the revised test methods may fail to detect a wide range of endocrine disruption effects. For every report that concludes that endocrine disrupting chemicals pose significant health risks, chemical manufacturers have commissioned scientists to craft rebuttals. Adding to the controversies, feminists have highlighted the ways in which endocrine disruption researchers often characterize the pernicious effects of endocrine disrupting chemicals by using problematic gendered language like “feminization” and “demasculinization.”

These controversies have prominent social, moral, and political dimensions. Many endocrine disruption researchers are deeply concerned about the public health impacts of failing to adequately test and regulate endocrine disrupting chemicals. Chemical manufacturers are likewise concerned to avoid burdensome regulations. Feminists are worried that the language used by endocrine disruption researchers might contribute to misogynistic social attitudes and problematic stereotypes of LGBTQ communities. Thus, parties with different social, moral, and political interests offer very different perspectives on how endocrine disruption research ought to be conducted and interpreted.

There is a philosophical view about science, prominent among mid 20th century philosophers of science, that scientific reasoning ought to be free from the influence of social, moral, and political values. Science is in the business of describing the way that the world *is*, while social, moral, and political values are about the way the world *ought* to be. Thus, according to "the value-free ideal," any influence of these values on scientific reasoning and practice represents a distortion of science.

Most philosophers of science no longer accept the value-free ideal. As early as 1953, Richard Rudner argued that ethical values are necessary for setting standards of evidence for accepting or rejecting a scientific hypothesis. In the late 20th century, feminist scientists and science studies scholars articulated the ways in which sexist assumptions shaped research in biology and social sciences and recommended interventions. In an influential work, Helen Longino (1990) argued that social beliefs and values were an ineliminable part of scientific reasoning, since these beliefs and values help to provide the background assumptions necessary to conduct scientific research. In the first decade of the 21st century, philosophers of science inspired by Heather Douglas (2000; 2009) have argued that scientists must make value-laden choices at every stage of scientific work.

Within the emerging tradition (pioneered by Rudner, Longino, Douglas, and others) of analyzing science as value-laden, this dissertation focuses on the roles of values in the descriptive choices and practices of endocrine disruption researchers. Chapter 2 provides an analysis of some of the criteria that govern descriptive practices in research into the

endocrine disrupting properties of the herbicide, atrazine. While not focusing directly on social, moral, and political values as do the subsequent chapters, Chapter 2 highlights the ways that descriptive goals organize atrazine research and necessitate the contributions of diverse scientific disciplines. Chapter 2 also introduces the concept of *criteria of characterizational adequacy*. These are criteria, endorsed by scientific communities, that set standards for scientific descriptions and the practices that generate them. These criteria help us to understand the contributions of diverse scientific disciplines to endocrine disruption research in general, and atrazine research in particular.

Chapter 3 analyses language choice in endocrine disruption research, specifically choices about whether or not to describe the effects of endocrine disrupting chemicals using gendered language. Some endocrine disruption researchers describe organisms impacted by endocrine disruptors as "demasculinized" and "feminized." Other researchers investigating similar phenomena eschew the use of this language. I argue that these language choices involve trade-offs among ethical values. Rudner (1953) and Douglas (2000; 2009) have focused on the roles of ethical values in managing "inductive risk." For Rudner and Douglas, inductive risk is the risk of being mistaken about the truth of a hypothesis. I use the case of gendered language in endocrine disruption research as an example to argue that there are forms of inductive risk that do not involve mistakes about the truth of hypotheses.

A major project of philosophers of science who reject the value-free ideal is providing an account of appropriate and inappropriate roles for social, moral, and political values in science. Traditional approaches to this project have focused on providing general norms for separating problematically value-laden science from science that is unproblematically value-laden. In Chapter 4, I argue that this approach is unlikely to meet with success. The diversity of scientific practices and aims suggest that any proposed demarcation of appropriate and inappropriate roles for values in science will probably not be applicable across all of science. I propose instead that philosophers focus on the local criteria of adequacy that scientists use to judge the quality of work in their own areas of expertise. Violations of these local criteria can serve as warnings that values may be playing an inappropriate role in scientific reasoning and practices. I focus on financial conflicts of interest in endocrine disruption research. Industry funding of research has been a significant topic of interest in recent science studies, and a major source of controversy in science policy discussions. Funding source has been found to be a strong predictor of study outcome in toxicology and pharmacology (Elliott and Resnik 2014). I provide a set of guidelines for inferring when violations of local criteria of adequacy are likely to be caused by financial conflicts of interest.

Chapter 5 investigates the roles of values in decisions about whether or not to be a "pluralist" about scientific terms. Pluralism about key scientific terms is now the default position in many philosophical and scientific debates. That is, philosophers and scientists often recognize that scientific terms have multiple useful meanings and definitions. But

in many scientific debates that are now pluralist in character, there were once warring factions of scientists and philosophers who insisted on the priority of their own definition of the scientific term in question. For example, species debates were once generally well-described as partisan fights over the correct definition of 'species.' Now many philosophers and scientists are more inclined to investigate the ways in which different species definitions allow for the pursuit of different scientific goals. In this way, the shift to pluralism has radically transformed the nature of debates where it has gained a foothold. This raises two questions: Which debates should we expect to become pluralistic in the future? And when is pluralism the appropriate orientation? I argue that in debates in which adopting a pluralist position would have consequences strongly disfavored by debate participants, we should not expect them to do so. Further, adopting a pluralist position can run counter to strong moral reasons endorsed by debate participants. In such cases, pluralism will arguably not be an appropriate position. I illustrate by analyzing debates about the terms 'potency' and 'endocrine disruptor' in toxicology. Debates about these terms are driven by both epistemic disagreements about standards for toxicological research, and ethical disagreements about trade-offs between regulatory burdens and the protection of public health and the environment.

Chapter 2: Localizing Criteria of Adequacy for Atrazine Research

2.1 Introduction

Philosophers have become skeptical about the prospects and motivation for theory reduction in biological and allied sciences. But participants in this skeptical consensus have historically not offered much in the way of well-developed alternative accounts of how various sciences and disciplines might be epistemically related (Rosenberg 1997; Robert 2004). In response to the decline of theory reduction, projects aimed at describing epistemic relationships among various biological and allied disciplines have been pursued via explanatory reductionist, anti-reductionist, and nonreductionist (often pluralist) strategies, but a need persists for detailed development of these strategies and application to particular case studies (Brigandt and Love 2012).

Here I develop and apply a nonreductionist problem-centered strategy for accounting for epistemic relationships among biological and allied disciplines. I use research into the endocrine disrupting effects of the herbicide, atrazine, as a case study. Atrazine researchers have claimed that answering pressing questions about the impact of atrazine on the environment and human health requires input from many disciplines.[1] Critical evaluation of these claims provides an opportunity for describing how certain disciplines

[1] I use "discipline" in a manner roughly synonymous with with Darden and Maull (1977)'s use of "field." A discipline has a core problem or set of problems (complex questions), a set of facts that have bearing on one or more core problems, and a set of explanations, descriptions, terms, concepts, models, and often hypotheses and theories related to one or more core problems. I follow Darden and Maull in not making principled distinctions between "supra-discipline," "discipline," and "sub-discipline." For a review of sociological methodology related to interdisciplinarity measurements, see Wagner *et al*. (2011)

play a role in offering answers to the complex questions that atrazine researchers seek to answer.[2] Articulating the roles played by the contributions of each discipline presents an opportunity to demonstrate how a nonreductionist epistemology can provide an account of disciplinary integration centered on solving particular problems and answering particular questions. Such an account is desirable because it promises to fill the void left by the abandonment of traditional theory reduction approaches for describing epistemic relationships among disciplines.

Love 2008's nonreductionist "problem agenda" account of local (as opposed to more broadly theoretical or unificatory) integration in the biological sciences deploys the concept of *criteria of explanatory adequacy*. These criteria, endorsed by particular epistemic communities and associated with particular problems and sets of problems, act as unifying constraints by specifying what sorts scientific explanations are adequate for the problems that motivate them.

But discussions of epistemic relationships among scientific disciplines should countenance the insight that explanation is not the only game in town. Solving scientific problems often requires a kind of characterization, achieved through various scientific practices, that is conceptually distinct from explanation (or, at least explanation in the

---

[2] I follow Love (2008)'s and Wagner *et al*. (2011, Table 1)'s distinction between "interdisciplinary" and "multidisciplinary." "Interdisciplinary" connotes a relatively stable integration of disciplines that may involve the creation of an "interfield theory" (Darden and Maull 1977) or new discipline. "Multidisciplinary" connotes research involving scientific problems that require more than one discipline, but that typically does not result in a stable integration or the formation of a new discipline; see also Bechtel (1993)'s discussion of "cross-disciplinary research clusters."

sense of providing causal mechanisms). Dose-response graphs in toxicology, for instance, are outcomes of experiments that have the proximal goal of giving a characterization of organismal responses to toxins rather than providing the mechanisms or processes generating those responses. Maps of concentrations of environmental pollutants are similarly an outcome of scientific experiments and modeling practices where the proximal goal is experimentally-grounded characterization of the concentration of pollutants at various spatial and temporal locations rather than the provision of an explanatory account of the distribution of the chemical in question, although the geographic pollutant concentration patterns so characterized may (or may not) later be an object of explanation.

Thus there is a need to augment nonreductionist problem-centered epistemologies of multidisciplinary integration with a treatment of the criteria by which various scientific disciplines might judge empirical characterization*s* (as opposed to explanations) and the processes by which such characterizations are generated to be adequate in the context of solving particular problems and sets of problems. I here give a treatment of local standards (*criteria of adequacy*) for concise descriptions (*characterizations*),[3] and show how these criteria play a role in multidisciplinary integration.

My central claim is that paying attention to *criteria of characterizational adequacy* allows philosophers pursuing nonreductionist problem-centered epistemologies to

[3] See Frigg (2006) for a description of the problem-space of scientific representation and its attendant literature, where the focus has tended towards providing a more global account of how models represent their target phenomena.

understand the activities of scientists, and the contributions of various disciples in multidisciplinary research, in a way that they could not if their epistemology countenanced only criteria of explanatory adequacy. Section 2.2 introduces atrazine research as a case study, highlights its multidisciplinary character, and introduces a puzzle about why a prominent atrazine researcher thinks that evolutionary biology is essential to atrazine research. Section 2.3 reviews and contextualizes the nonreductionist strategy of Brigandt and Love (2012), and describes their central concepts of *problem agendas* and *criteria of explanatory adequacy*. Section 2.4 introduces the idea that some complex scientific questions can be shared among problem agendas, discusses some of the resulting implications, and introduces examples of such shared questions in atrazine research. Section 2.5 introduces the problem agenda of environmental toxicity and the concept of criteria of characterizational adequacy, shows how research in the environmental toxicity problem agenda is constrained by criteria of explanatory and characterizational adequacy, and addresses two objections. Section 2.6 examines the problem agenda of developmental endocrine function, and highlights two of its important criteria of explanatory adequacy. Section 2.7 shows how some of the criteria of explanatory and characterizational adequacy that constrain solutions to problems on the environmental toxicology and developmental endocrine function problem agendas also constrain answering narrower questions germane to assessing atrazine impacts on amphibians. Section 2.8 provides a solution to the puzzle of the contributions of evolutionary biology to atrazine research; models provided by evolutionary biology help to fulfill criteria of explanatory and characterizational adequacy associated with atrazine

research. Understanding these contributions requires attention to both explanatory and characterizational criteria.

2.2. Atrazine research as case study

Atrazine is a top selling herbicide that is a highly persistent and widely distributed ground and surface water pollutant (Jablonowski *et al*. 2011; Thurman and Cromwell 2000). The effects of atrazine on amphibians and the contribution of these effects to global amphibian decline have been the subject of much research, requiring the input of many disciplines. Work in molecular biology, biochemistry, developmental biology, endocrinology, physiology, and organismal biology has supported the view that atrazine acts as an endocrine (hormone) disruptor in vertebrate organisms; according to a candidate mechanism, it induces a class of enzymes (aromatases) that convert androgens (e.g., testosterone) into estrogens (e.g., estradiol) (Hayes *et al*. 2011). Atrazine exposure is now widely believed to have diverse effects on many different kinds of vertebrate organisms with respect to gene expression, endocrine function, sexual development, predator avoidance behavior, reproductive success, olfaction, and immune system function (Jin *et al*. 2014; *Santos et al*. 2012; Hayes *et al*. 2011; Hayes 2005; Rohr and McCoy 2010a).[4,5]

---

[4] However, many of these conclusions are disputed by studies and researchers funded by atrazine's manufacturer, Syngenta Crop Protection LLC (Van Der Kraak *et al*. 2014; Solomon *et al*. 2013; 2008).
[5] In 2012, the USEPA concluded that atrazine "does not consistently affect amphibian gonadal development" (USEPA 2012, 62). However, the agency appears to have reached this conclusion on the basis of a single study funded by Syngenta Crop Protection LLC, Kloas *et al*. (2009). See Boone *et al*. (2014) for an account of why other studies were excluded. Thus, the USEPA's regulatory decision here is

Describing and predicting atrazine persistence, transport, and exposure has involved input from diverse disciplines including hydrology, agricultural science, geology, soil science, environmental chemistry, and meteorology (Hayes *et al* 2011, Rohr and McCoy 2010a, Hayes 2005). Tyrone Hayes, a leading researcher on atrazine's endocrine disrupting effects on frogs, claims that,

> To truly assess the impact of atrazine on amphibians in the wild, diverse fields of study including endocrinology, developmental biology, molecular biology, cellular biology, ecology, and evolutionary biology need to be invoked. To understand fully the long-term impacts on the environment, meteorology, geology, hydrology, chemistry, statistics, mathematics and other disciplines well outside of biology are required. (2005, 321)

Although understanding physiological developmental mechanisms seems key to understanding abnormal amphibian development resulting from exposure to endocrine disruptors like atrazine, and research on atrazine transport and persistence seems obviously necessary to infer exposure rates and magnitudes, it is not immediately clear what it is about this question that requires input from other disciplines, e.g. evolutionary biology. What justifies Hayes' claim that evolutionary biology is *required*? A framework for structuring multidisciplinary inputs within the atrazine research program can help us articulate the roles played by various disciplines in answering the question of the impact of atrazine's endocrine disrupting effects on amphibians in the wild and

---

consistent with the view that studies finding that atrazine has no significant impact on amphibian development are predominantly funded by atrazine's manufacturer (Hayes 2004).

thereby allow us to critically evaluate claims (like Hayes's) for the necessity of particular disciplines.

Love (2008) develops an account of localized integration in the sciences based on what he calls "problem agendas," or sets of problems (complex questions composed of simpler questions) related to a particular epistemic goal. Here I cast the impact of atrazine's endocrine disrupting effects on amphibians as a simpler question within the problem (more complex question) of the impact of anthropogenic endocrine disruptors on the environment. I will interpret environmental endocrine disruption as a problem shared by the problem agendas of environmental toxicity and developmental endocrine function. To clarify this interpretation, I will describe Love's notion of *criteria of explanatory adequacy*, the criteria by which explanatory answers to problems (complex questions) on a particular problem agenda are judged to be adequate or inadequate by the epistemic communities working on the agenda. I will then introduce the complementary concept of *criteria of characterizational adequac*y (CCA), criteria by which empirically grounded characterizations and the practices by which they are generated are judged by research communities to be adequate or inadequate with respect to particular epistemic goals.[6]

[6] One might wonder how criteria of adequacy relate to the concepts and conclusions found in the extensive literature on modeling and confirmation (e.g., Sprites *et al*. 2001; Forster and Sober 1994). While we can obviously interpret, e.g., dose-response graphs as models that are subject to evaluation by and the outcome of various forms of statistical and probabilistic analyses that can putatively demonstrate varying degrees causality, these are not the only possible philosophically interesting interpretative frameworks for such graphs. Here I am interested in interpreting such graphs (qua characterizations, or concise descriptions) in terms of a set of criteria that members of the atrazine research community place on them, showing how these criteria contribute to the multidisciplinary nature of atrazine research, and describing some roles that these criteria play in disagreement within that community. I do not take the criteria of adequacy that I treat here to be exhaustive. For example, I do not discuss arguments about the relative virtues of linear versus nonlinear (or montonic versus nonmonotonic) families of curves for describing dose-response relationships

I will show in section 2.7 how the criteria of characterizational and explanatory adequacy of the two problem agendas of environmental toxicity and developmental endocrine function structure disciplinary inputs with respect to the narrower question of the impact of atrazine's endocrine disrupting effects on amphibians. I will then show in section 2.8 how a set of criteria of explanatory and characterizational adequacy drawn from the two problem agendas can make clearer the contributions of evolutionary biology to the question of the impacts of atrazine on amphibians in the wild.

2.3 Nonreductionism, problem agendas, and criteria of explanatory adequacy

Contra more radically permissive pluralist accounts (e.g., Dupré 1993), advocates of the so-called "pluralist stance" have contended that the nature of the specific scientific problem or question being addressed constrains the "variety of acceptable classificatory or explanatory schemes." (Kellert *et al*. 2006) [7] Taking onboard this feature of the pluralist stance, Love (2008) and Brigandt (2010) have offered structured accounts of local integrations in evolutionary developmental biology (evo-devo) that focus on solving particular problems and explaining particular explananda. These local integrations need not, for the authors, necessarily be part of any broader unificatory theoretical reduction (Nagel 1961; Schaffner 1993) or unificatory explanatory ideal (Kitcher 2001). Love and

---

in endocrine disruption research, and the role that, e.g., the Akaike information criterion (*qua* proposed criterion of adequacy on choosing among families of curves) might play in that debate. See Elliot (2011, Ch. 2) and Forster and Sober (1994), respectively, for an introduction to these issues.

[7] See also Frigg (2006, section 6).

Brigandt's views do, however, emphasize the important role of more problem-specific explanatory (as opposed to theoretical) reductions in biological explanation.

Where theory reduction approaches contend that laws describing "lower" mereological levels are more fundamental in explanation, on the problem-centered view, explanatory fundamentality "*varies with the specific problem at hand*." (Brigandt 2010) Thus, Brigandt and Love's problem-centered integrative frameworks are *nonreductionist* in that they do not necessarily ascribe explanatory fundamentality to lower level epistemic units (laws, theories, mechanisms, models, etc.).[8] However, these frameworks are not *antireductionist* because they reserve a place for reductive explanation when such explanation is called for by the nature of the specific scientific problem or problems under consideration. In this regard, Love and Brigandt's views are compatible with views that reductive biological explanations are appropriate in particular circumstances (Waters 1990; Sober 1999).

Love (2008) characterizes *problem agendas* as sets of problems (complex questions) related to a particular complex epistemic goal. Problem agendas are united in part by *criteria of explanatory adequacy*, criteria for judging the acceptability of candidate solutions to the problems composing the agenda (875).[9,10] Similarly, Brigandt (2010)

---

[8] See Craver (2007, Chapter 5) for a critical treatment of levels.

[9] See also Plutynski (2005) for arguments to the conclusion that formal analyses in the sciences serve to "delineate the conditions of adequacy of any explanatory story for some domain."

[10] Talk of problem agendas naturally raises several questions related to values in science discussions, especially when considering controversial cases like atrazine research. I will not here address questions about demarcating epistemically and ethically permissible problem agendas from those that are not, other

refers to sets of related scientific explananda as a *complex explanandum*. Against theorists who argue for (typically reductive) stable theoretical integration or unification of diverse fields of science (Nagel 1961; Schaffner 1993), both Brigandt and Love (as well as Wimsatt 1974) argue that integration of multiple fields of study can profitably be localized to particular epistemic goals without necessarily requiring more global theoretical integration or unification.

Criteria of explanatory adequacy are central to Love's account of localized integration. Such criteria make possible "an explicit account of how different areas of research make their contribution without one being more fundamental than another." (2008, 875) Because calls for multidisciplinary research typically arise out of the need to solve problems and answer questions rather than a need for theory-building or testing, what is needed is an account of what ought to count as adequate answers to the complex questions driving the research.

Love uses the problem agenda of evolutionary innovation and novelty as an example to illustrate the concepts of problem agendas and criteria of explanatory adequacy. Problems on the innovation and novelty agenda include, e.g., "How did vertebrate jaws originate?" and "How did avian flight originate?" Although perhaps superficially resembling more ordinary questions (e.g., Who broke the window?) these problems are not standard interrogatives of the sort that can be answered with a single proposition.

---

than to say that the complex questions that compose the agendas are evaluable in terms of, and motivated by both ethical and epistemic values. I address such concerns in Chapter 3.

These problems, due to their complexity and the diversity of simpler questions that they naturally engender, are thought to require multidisciplinary input from developmental, evolutionary, molecular, and systematic biology (2008, 879).

Love claims that the inputs of these disciplines can be structured by the criteria of explanatory adequacy associated with the project. For Love, adequate explanations of the origination of radical evolutionary changes in phenotype must meet three criteria grounded in the nature of the explananda. First, the explanation must address both form and function; e.g., explanations of the origination of vertebrate jaws must include considerations related to how these sorts of jaws function given the particular forms that they take. Second, accounts of origination must explain innovation and novelty at all biological levels of organization as well as relations among these levels, e.g., genetic, cellular, modular, organismal, and population levels (Love 2008). And finally, there is the third criterion of "degree of generalization," which deals with how different problems within the agenda are related. For the case of evolutionary novelty, this criterion can be broken into two further questions. 1- "Can investigations of particular novelties be generalized to other research on different innovations or novelties?" and 2- "Can investigations of model systems be generalized to the phylogenetic juncture relevant to the innovation or novelty under scrutiny?" (Love 2008) The concern addressed by this criterion of explanatory adequacy is the appropriateness of generalizations from one problem or question within the agenda to others. [11]

[11] A reviewer has raised the question of how criteria of adequacy are to be justified. In general, criteria of adequacy are justified by their capacity to promote the values (both epistemic and non-epistemic) that

Table 1 below provides examples from Love's problem-agenda account of evolutionary innovation and novelty.

---

motivate the research that the criteria govern. In this essay, I will not provide a more detailed account of the justification of criteria of adequacy, but will rather take them as given by the relevant scientific communities, and address the separate question of how these criteria organize the research that they govern. These criteria emerge from historical dialectics between and among scientists and science critics about how to best achieve the epistemic and non-epistemic goals that motivate the research in question. Because of this historical dimension, criteria of adequacy may change throughout the history of a problem agenda, and so any particular set of criteria of adequacy are usually not tied in a necessary way to a particular problem agenda. It is also often the case that criteria of adequacy are contested (Brigandt 2015). I address the justification of, and values trade-offs with respect to, criteria of adequacy in Chapter 3.

Table 1	
Problem Agenda	**Evolutionary Innovation and Novelty***
Focus of Problem Agenda	Explanations of evolutionary innovations and novelties*
Criteria of Explanatory Adequacy	Explanations must address both form and function*
Problems (complex interrogatives)	How did the vertebrate jaw arise?*
Questions (simpler interrogatives)	What mechanisms currently account for vertebrate jaw development?*

*(Love 2008)

2.4 Shared problems and questions in environmental research

Environmental problems, due to their inherent complexity, are exemplary of the sorts of problems that require multidisciplinary input for generating adequate solutions (Love 2008, 875). The question of atrazine's effects on amphibians in the wild as a result of its endocrine disrupting properties can be viewed as a simpler question located within the environmental problem (complex question) of endocrine disruptors and their ecological impacts.[12] This problem is shared by the problem agendas of environmental toxicity and developmental endocrine function, each with its own criteria of explanatory and characterizational adequacy. These criteria will be shown to constrain and unify attempts at answering questions clustered around the impact of atrazine on amphibians.

To illustrate this, I will begin by offering some plausible sample questions germane to the broader question of atrazine's role as an environmental amphibian endocrine disruptor. Notice that the levels of biological organization at which the questions are aimed increases sequentially. The first question is aimed at the biochemical and genetic levels; the second is aimed at the morphological level; the third is aimed at the population level, and the fourth is aimed at global scale ecological phenomena and impacts on higher-level

[12] See Krimsky (2001) for arguments to the conclusion that "the endocrine disruptor thesis" does not function as a theory. For Krimsky, this is a short-coming to be overcome. See Love (2014) for arguments to the conclusion that not all science needs to be organized by a central theory or theories.

taxa. The species named in the first through the third question are reflective of some of the organisms that are frequently used in such research (Hayes 2005; 2011).

1. What effect does atrazine exposure at a given concentration and duration have on CYP19 (aromatase gene) expression in *Xenopus laevis?*
2. How do the morphological effects of given concentration and duration of atrazine exposure in *Hyperolius argus* differ depending on the developmental stage at which exposure occurs?
3. What impacts do atrazine's endocrine disrupting effect have on *Rana pipiens* populations in Midwestern corn growing regions?
4. Do atrazine's endocrine disrupting effects play a significant role in global amphibian decline?

I want to suggest that answers to these and similar and related questions will be constrained by criteria of explanatory and characterizational adequacy drawn from the two problem agendas in which problem of environmental endocrine disruption and the question of atrazine's impact on amphibians resides. But first, I need to sketch these problem agendas, environmental toxicity and developmental endocrine function. In discussing this second problem agenda of environmental toxicity, I will introduce a distinction between explanation and empirical characterization and make explicit how the concepts of criteria of explanatory adequacy and criteria of characterizational adequacy are analogous.

2.5 The environmental toxicity problem agenda and criteria of characterizational adequacy

Environmental toxicology has been described as "the study of the impacts of pollutants on the structure and function of ecological systems." (Landis *et al*. 2010, 1) Its focus is the identification of (primarily anthropogenic) toxic agents and the establishment of the causal bases of their toxicity (Landis *et al*. 2010, Chapter 3).[13] These two epistemic goals highlight a distinction between empirical characterization and explanation. In the case of identifying toxic agents, the goal is identifying and characterizing the effects of a chemical and classifying it according to its toxic properties, a task of description and evaluation (characterization). In the case of identifying causal bases of toxicity, the goal is explanatory, concerned with providing a causal account of the processes by which a chemical gives rise to toxic effects. Such explanatory goals seem clearly amenable to constraint by criteria of explanatory adequacy as Love develops the concept. E.g., an explanation of the mechanism by which atrazine is toxic to plants, starvation and harmful oxidative effects due to interruption of plastoquinone-binding in photosystem II (Appleby *et al*. 2001), is constrained by the criterion of explaining higher-level physiological effects by reference to lower-level biochemical processes.[14]

---

[13] Although Landis *et al.'s* description is of environmental toxicology as a discipline, one may also apply this description to the problem agenda of environmental toxicity.

[14] Such constraints are suggested in the Hill criteria, *plausibility* and *coherence* (Hill 1965; Hayes *et al*. 2011).

It seems strange, however, to say that the descriptive and evaluative goals of describing and classifying chemicals and their impacts according to toxicity are constrained by *sensu stricto* criteria of *explanatory* adequacy. After all, the goal is description and classification rather than explanation (although as we will see, some descriptive and classificatory claims derive their inferential justification from explanatory accounts).[15, 16] Rather, such attempts at scientific characterization are constrained by *criteria of characterizational adequacy*. Criteria of characterizational adequacy (CCA) are constraints on empirically grounded characterizations that specify what counts as adequate justification for those sorts of characterizations.[17]

Why do we need to recognize this additional sort of criteria of adequacy? Waters (2007) points out that the findings of so-called "exploratory" experiments can have significance for various scientific goals other than explanation and theory development, including knowledge about experimental manipulation and conceptual development to guide future

---

[15] Lewontin (1974, Ch. 1) draws a similar distinction between descriptive and dynamic sufficiency.

[16] Further, the requirement that classificatory or descriptive claim be grounded in an explanatory or theoretical framework can itself function as a kind of criterion of characterizational adequacy. See, e.g., Rothwell (2010) for arguments to the conclusion that behavior coding in focus group research must be grounded in group dynamic or group process theory.

[17] A brief note on a possible mapping of my terms, "characterization" and "criteria of characterizational adequacy" to the influential terminology of Bogen and Woodward (1988): As I see it, it is possible to have characterizations of both data and phenomena, as well as attendant criteria of characterizational adequacy for characterizations of both data and phenomena. It is defensible in some contexts to think of both "data" and "phenomena" as relational concepts such that the same characterization can stand in a data relationship to some phenomenon and in a phenomena relationship to some set of data. For example, a dose-response graph can characterize the phenomenon of the response of a population of frogs to a putative endocrine disrupting chemical, or the same graph can be taken as a datum in relation to (i.e.,"evidence for the existence of") some other putative phenomena, e.g., the detrimental effect of endocrine disrupting chemicals on global amphibian populations. With regard to questions about the extent to which phenomena are constructed, I am sympathetic to the view that phenomena are neither given to us ready-made by nature nor merely stipulative (Massimi 2011).

research. Minimally, explanation requires explananda, and those explananda often require scientific investigations to in order to be recognized as things wanting explanation and to disclose the ways in which they might be experimentally manipulated or exploited in the future (O'Malley 2007; Burian 1997; Steinle 1997). These philosophical treatments of exploratory experiments remind us that the explanation-focused activities of theory development and testing do not exhaust the set of philosophically interesting scientific activities.[18,19]

Waters (2007) discusses some of the ways that exploratory experiments are pursued with respect to goals other than theory development (even though such experiments are often "theory-informed"). Certain "fact-gathering" activities (to use the language of Kuhn 1970) appear to be engaged in attempts at scientific characterization rather than explanation, e.g., the discovery of new entities or previously uncharacterized properties of entities (Waters 2007). Metagenomics, for instance, focuses on disclosing patterns in large amounts of sequenced DNA "in order to explore currently uncharacterized microbial entities and processes" (Waters 2007; O'Malley 2007).

An exclusive focus on explanation at the expense of characterization occludes the importance of such non-explanatory methods and disciplines and the ways that constraints on characterization structure experiments and research programs. Love

[18] See also Hacking (1983) for a critique of theory-centrism in philosophy of science.
[19] See also Elliot and McKaughan (2009)'s arguments that the outcomes of testing activities in the "context of justification" are not independent of the values present in the "context of discovery."

(2008) has shown how explanatory accounts within problem agendas are constrained and structured by criteria of explanatory adequacy. Recognizing the importance of non-explanatory methods and disciplines within various problem agendas requires an analogous account of scientific characterizations in terms of CCA.

To illustrate how the concept of CCA operates, consider the case of dose-response graphs common to the problem agenda of environmental toxicology. A dose response graph is "a graph describing the response of an enzyme, organism, population, or biological community to a range of concentrations of a xenobiotic." (Landis *et al*. 2010, 36) The task here is characterizational rather than explanatory; such graphs have no necessary reference to causal mechanisms explaining the phenomena represented by the graph. However, the production of such a characterization is constrained by certain criteria. For example, the points on the graph must make reference to a concentration of the xenobiotic and must be compared to a control in which the xenobiotic is absent, i.e., the "normal" behavior of the enzyme, organism, population, or biological community under consideration. These *concentration relative* and *compared to control* CCA allow us to see the contributions of exploratory research aimed at characterizing the properties of entities in a way that we could not if we considered only criteria of explanatory adequacy.[20]

---

[20] See Hill (1965)'s treatment of "biological gradient." Of course not all studies will have a control group that is completely free of the putative toxin. Sometimes it is sufficient for the control group to be exposed to the compound of interest at non-zero levels, but below a threshold dose (a dose at which there are no apparent statistically significant effects). See Landis *et al*. (2003, 40) for an introduction to how threshold doses are established.

To see evidence of these criteria in science criticism, consider the following use of a *compared to control* criterion. Coady *et al*. (2004) exposed larval frogs to nominal concentrations of atrazine of 0 (control), 10, and 25 μg/L. The stated goal of the study was characterizational, to "to determine the response of larval *Rana clamitans* to atrazine by assessing metamorphosis and reproductive indices in animals exposed during larval development." (942) In order to determine that actual levels of atrazine exposure matched nominal levels (and thus provide evidence of meeting *compared to control* and *concentration relative* criteria of adequacy), tank water from the treatment groups was analyzed using immunosorbent assay and gas chromatography-mass spectroscopy. These analyses were interpreted as showing that the atrazine concentration in the control group was between 0.07 and 0.25 μg/L (Coady *et al*. 2004, Table 1). Two review articles have described Coady *et al*. (2004) as showing that atrazine has no effect on several developmental endpoints (Van Der Kraak et al. 2014; Solomon et al. 2008). In contrast, Rohr and McCoy (2010b) claim that because the controls in Coady *et al*. (2004) were contaminated with atrazine above a plausible threshold dose (Hayes 2004), such claims of no effect are specious. Rather than saying that there was no effect of the atrazine treatments with respect to these measurements, we should say that no such conclusions can be drawn from the dose-response graphs because the putative controls contained atrazine concentrations above a plausible threshold dose for the effects under investigation (Rohr and McCoy 2010b). In short, because a *compared to control* criterion of adequacy was not met in the study, the dose-response graphs do not well-

characterize the effects of atrazine on the frogs in the study as compared to an atrazine-free control, and the findings of no effect are not well-supported.

Such criticisms highlight the importance of precise characterizations of concentrations in toxicology. Indeed, much of the research activity in the environmental toxicity problem agenda is aimed at characterizing concentrations (in cells, organs, organisms, particular habitats, *etc*.) (Rohr and McCoy 2010a, Hayes *et al*. 2011). Properties of entities at characterized concentrations must then be compared to properties of entities free from the putative toxin, and the characterization of these toxin-free properties requires research. In the case of atrazine, the near ubiquity of the chemical in fresh water supplies, and its potential for effects at very low doses has necessitated the development of sophisticated filtering techniques, careful attention to laboratory hygiene, and a variety of chemical analytic techniques in order to characterize the properties of biological entities in their atrazine-free conditions. Additionally, due again to atrazine's near ubiquity in the environment, the *compared to control* criterion has made essential early characterizations of frog morphology in the wild (e.g. Witschi 1929), characterizations made before the wide-spread application of atrazine began in the 1950s (Hayes *et al*. 2011; Hayes 2005; Hayes 2004; Rohr and McCoy 2010a).

I will note here that this sort of empirical characterization is distinct from the sort of description involved in so-called "mechanistic" explanation. Mechanistic explanation is often thought of as a description of the entities and activities that compose a mechanism

producing a phenomenon (Tabery 2004). The empirical characterization treated here differs from the descriptions of entities and activities in mechanistic explanation in that empirical characterization is not necessarily concerned with providing an account of a mechanism or an explanation of a phenomenon. Empirical characterization, as in the cases of dose-response curves and pollutant concentration maps, is often performed with other goals in mind, e.g., describing patterns for future explanation, modeling, or prediction, or developing new experimental techniques. For instance, where an empirical characterization of the response of a biological entity to an environmental pollutant typically takes the form of a dose-response graph in an effort to provide a useful representation or concise description of the phenomenon (for, e.g., prediction, future experimental manipulation, or future explanation), a mechanical explanation of that response would typically invoke and describe (often mereologically "lower") entities and activities, i.e. mechanisms of toxicity, in order to give a causal account.

It might be objected that the concept of criteria of characterizational adequacy is superfluous. For example, Woodward's (2006) manipulation account of causation stipulates epistemic conditions for establishing causation in nature. Perhaps we can understand criteria like *compared to control* and *concentration relative* in terms of the criteria governing a manipulation account of causation, and so there is not a need to posit additional criteria or establish a new term for those criteria. I do not doubt that manipulist (and other) accounts of causation can provide justifications for criteria like *compared to control* and *concentration relative*. However, this objection misunderstands the goal of

introducing criteria of characterizational adequacy in two ways. First, problem-centered nonreductionist epistemologies are aimed at understanding scientific practice in terms of the criteria endorsed by the scientists and used in their criticisms and discourses (Brigandt and Love 2012). While criteria like compared to control may be justified in terms of their ability to help establish Woodwardian manipulist causation, such considerations do not appear in scientific discourses about atrazine. Second, criteria of characterizational adequacy are criteria governing the production and justification of concise descriptions in science. Thus, criteria of characterizational adequacy govern not only graphs, but also other empirically-grounded characterizations including names (see Chapter 2), maps, diagrams, *etc*. We need the concept of criteria of characterizational adequacy, in contrast to criteria of explanatory adequacy, to talk about the criteria, endorsed by scientists and used in their criticisms, that govern these characterizations in the context of providing nonreductionist problem-centered epistemologies of multidisciplinary research.

Table 2 below provides a comparison of examples from Love's problem-agenda account of evolutionary innovation and novelty and my problem-agenda account of environmental toxicity.

Table 2		
Problem Agenda	**Evolutionary Innovation and Novelty ***	**Environmental Toxicity**
Focus of Problem Agenda	Explanations of evolutionary innovations and novelties*	Characterization and explanation of toxic effects of anthropogenic chemicals in the environment
Criteria of Adequacy (Explanatory or Characterizational)	Explanations must address both form and function*	Characterizations of effects must make reference to toxin concentration (in organisms or the environment)
Problems (complex interrogatives)	How did the vertebrate jaw arise?*	What effects do anthropogenic endocrine disruptors have on organisms in the environment?
Questions (simpler interrogatives)	What mechanisms currently account for vertebrate jaw development?*	What impacts do triazine herbicides have on amphibians in the wild?

*Love (2008)

I wish to briefly address another objection. A philosopher might be tempted to reconstruct the epistemic function of dose-response graphs in ecotoxicology and their associated criteria of adequacy exclusively in terms of the role they play in hypothesis testing, and thus in evaluating candidate explanations. What I have called criteria of characterizational adequacy, would then, in this sense be criteria for judging the acceptability of candidate explanations, or criteria of explanatory adequacy (as developed by Brigandt and Love 2012). Maybe we should say that toxicologists generating dose-response graphs are (necessarily) testing a hypothesis of whether a chemical has some particular effect on some particular biological endpoint. No one would deny that dose-response graphs sometimes have a hypothesis testing function (*qua* evidence). But I deny that this is the only philosophically interesting role for such graphs and their criteria of adequacy, or the sole motivation for their creation. Although a toxicologist may have specific hypotheses about specific biological endpoints and effects in mind when designing and carrying out an experiment, in other cases, a defensible and plausible interpretation is that that the experimenter designs an intervention on a biological system (i.e., the introduction of a putative toxin) with the goal of characterizing the state of the intervened upon system relative to some control.[21] (This interpretation is consistent with, for example, Coady et al. (2004)'s atrazine study measuring many developmental endpoints rather than a single targeted endpoint or set of endpoints.) The proximal goal is to characterize the result of the intervention rather than to test some particular hypothesis.

[21] See Leonelli (2009) for arguments for the portability of data.

This practice is analogous to the practice, endogenous to genetics and developmental biology, of characterizing the phenotype of a mutant (with the control group in toxicology being the analogue of *wild-type* organisms in genetics and development research). To be sure, such methods are often described *post hoc* by the scientists themselves as hypothesis testing, but this may reflect the demands of funding agency guidelines rather than an honest and reflective appraisal of epistemic methodology (O'Malley *et al*. 2009).

Continuing, we have seen here that the problem agenda of environmental toxicity has at least three important criteria of explanatory and characterizational adequacy. First, explanations of toxicity must appeal to mechanisms at all the relevant levels of biochemical and biological organization.[22] Second, characterizations of toxic effects must make reference to toxin concentration and, third, be compared to toxin-free (or below threshold) controls. Because a full account of environmental toxicity must contain both empirical characterization and explanation, the problem agenda as a whole is constrained by both kinds of criteria. Characterizations and criteria of adequacy may have roles to play in testing specific hypotheses in some contexts, but this does not exhaust their uses or the motivations for their creation.

2.6 The developmental endocrine function problem agenda

[22] Which conception of levels (see Craver 2007) is used, and which are relevant will depend on the questions being asked. See Solomon *et al*. (2008, 763) and Hayes *et al*. (2011) for evidence that such level criteria are actual criteria of adequacy at work in ecotoxicology.

The purpose of the study of developmental endocrine function is to provide an account of the biochemical processes and pathways of hormone synthesis, storage, and physiological function during organismal development. Problems (complex questions) comprising a developmental endocrine function problem agenda include "how do sex steroids control development?" and "how do thyroid hormones control development?" Solutions to these sorts of problems are constrained by the need to address causality at multiple levels of biochemical and biological organization and the need to justify generalizations from insights about pathways and processes in model organisms to claims about other organisms (roughly, Love's second and third criteria of explanatory adequacy) (Chester-Jones *et al*. 2013; Love 2008, 880-881).

Research into sex steroid determination of sexual development provides examples of these criteria in action. Comparative endocrinology research has discovered that androgens and estrogens control sexual development across the vertebrates, although the developmental effects of these hormones vary by taxa, imposing limits on generalizations made across taxa (Chester-Jones *et al*. 2013). The effects of these hormones tend to be "organizational" and irreversible at earlier stages of development and "activational" and reversible in adults.[23] Explanations of these developmental effects (and their relative permanence) make reference to biochemical pathways, gene expression, cellular metabolism and differentiation, and tissue development (Chester-Jones *et al*. 2013).

[23] See Wimsatt and Schank (2004) on "generative entrenchment."

We can see reference to *multi-level mechanism* and *generalization* criteria of explanatory adequacy at work in the discourses of atrazine researchers with respect to the developmental effects of atrazine (*qua* endocrine disruptor). For example, Solomon et al. (2008) write that "Atrazine has been proposed to exert adverse effects on the reproductive fitness of animals including mammals, fish, and amphibians… mechanisms observed in one species are often uncritically cited as support of proposed mechanisms in other species[.]" (p.739) Here, the authors are complaining of a failure to meet a *generalization* criterion of explanatory adequacy in some studies. They are also skeptical of the plausibility of various proposed mechanisms by which the biochemical action of atrazine can lead to developmental changes at higher levels of organization and thus impact reproductive fitness (739-748). In reply, Hayes *et al*. (2011) aim to present 1) "experimental evidence that the effects of atrazine on male development are consistent across all vertebrate classes examined" and 2) "summary of the mechanisms by which atrazine acts as an endocrine disruptor to produce these effects" (p. 64). The first is aimed at fulfilling a *generalization* criterion of explanatory adequacy while the second is aimed at fulfilling a *multi-level mechanism* criterion of explanatory adequacy.

Table 3 below provides a comparison of examples from Love's problem-agenda account of evolutionary innovation and novelty and my problem-agenda accounts of environmental toxicity and developmental endocrine function.

Table 3

Problem Agenda	**Evolutionary Innovation and Novelty ***	**Environmental Toxicity**	**Developmental Endocrine Function**
Focus of Problem Agenda	Explanations of evolutionary innovations and novelties*	Characterization and explanation of toxic effects of anthropogenic chemicals in the environment	Explaining roles of hormones in organismal development
Criteria of Adequacy (Explanatory or Characterizational)	Explanations must address both form and function*	Characterizations of effects must make reference to toxin concentration (in organisms or the environment)	Explanations must make reference to all relevant levels of biological organization
Problems (complex interrogatives)	How did the vertebrate jaw arise?*	What effects do anthropogenic endocrine disruptors have on organisms in the environment?	How do sex steroids control organismal development?
Questions (simpler interrogatives)	What mechanisms currently account for vertebrate jaw development?*	What impacts do triazine herbicides have on amphibians in the wild?	How do testosterone and estradiol control sexual development in amphibians?

*Love (2008)

2.7 Criteria of adequacy for atrazine research

Now I wish to show how some of the criteria of explanatory and characterizational adequacy that constrain solutions to problems on the environmental toxicology and developmental endocrinology problem agendas also constrain answering narrower questions germane to assessing atrazine impacts on amphibians. First, because environmental toxicity characterizations must make reference to controls free from the putative toxin (or below some plausible threshold dose), answers to the question of the impacts of atrazine must be predicated on atrazine exposure effects compared to atrazine-free controls or hypothetical populations (a criterion of characterizational adequacy). Much of the important research in the "emerging" science of amphibian endocrine disruption has been made possible by basic research on, e.g., CYP19 gene expression, aromatase catalysation of estrogenesis, sex steroid control of sexual differentiation during amphibian development, amphibian reproductive anatomy and behavior, and population genetic modeling of amphibian evolution (2005, Hayes *et al* 2011) Such studies provide controls against which the effects of atrazine at environmental concentrations inferred by sampling (as well as transport and persistence studies) can be compared. This criterion also provides grounds for the rejection of some proposed answers to questions about atrazine's effects on amphibians. Some authors, for instance, have proposed that hermaphrodism is widespread in wild amphibian populations in the absence of atrazine exposure (Carr *et al*. 2003). However, this conclusion was based on field and laboratory

studies in which the controls are thought to have been exposed to environmental atrazine, possible at relatively high concentrations (Hayes 2004; 2005, Rohr and McCoy 2010a)

Second (similar to the second of Love's criteria for explanations of innovation and novelty), adequate answers to questions about atrazine's effects on amphibians must give an account of all the relevant levels of biological organization (a criterion of explanatory adequacy). For instance, an answer to the third question in the list of questions given in section 4 (p. 14) would plausibly give a causal account of the effects of atrazine on Midwestern leopard frog populations by invoking atrazine's role in inducing aromatase gene expression, enhanced rates of estrogenesis in developing male frogs, "demasculization" and "feminization" of affected individuals, decreased reproductive success, and, finally, population level outcomes, e.g. local extinction or adaptation.[24] The absence of this sort of relatively complete causal chain providing mechanisms at multiple levels would imply "black boxes" that would potentially frustrate attempts to explain higher level phenomena in terms of atrazine exposure. It is here that the nonreductionist (as opposed to anti-reductionist) nature of my account comes to the fore. Because research into the endocrine disrupting properties of atrazine includes questions about how lower level molecular mechanisms give rise to higher level phenotypic and population level effects, a reductionist criterion of adequacy is appropriate. But the propriety of the reductionist criterion is grounded in the nature of the question being asked rather than an *a priori* commitment to reductionist explanation.

[24] See Chapter 2 for a treatment of the use of gendered language in endocrine disruption research.

Third, (similar to the third of Love's criteria for explanations of innovation and novelty), adequate answers to questions about atrazine's endocrine disrupting effects on amphibians in the wild must be constrained by considerations of generalization. With respect to the meeting the *generalization* criterion in their review of atrazine research, Hayes *et al*. argue that,

> Androgens are necessary for testicular development and maintenance of male germ cells in all vertebrates. Thus, given that atrazine reduces androgen production and stability, it is reasonable to expect the demasculinization effect in all vertebrates. On the other hand, partial or complete sex reversal of gonads by estrogens is limited to fish and amphibians, which lack morphological distinguishable sex chromosomes, and to reptiles with environmental sex determination, which lack sex chromosomes altogether. Birds and mammals, which have genetic sex determination and highly differentiated sex chromosomes are not susceptible to estrogen-induced sex reversal of the gonads. As such, while depleting androgens will impair testicular development and induce testicular lesions (such as the effects described here) in all vertebrates, increasing estrogen production (via atrazine) would not be expected to induce feminization of the gonads in birds and mammals, but would do so in fish, amphibians, and reptiles with environmental sex determination. (70)

There seem to be two dimensions of generalization at play here. The first concerns inferring the presence of mechanisms of endocrine disruption (e.g., changes in sex steroid ration due to aromatase induction) in a given clade or clades from the presence of such mechanisms in another clade or clades. The second concerns generalizing from the (biochemical, cellular, organismal, or populational) effects of endocrine disruption in one clade to similar effects in another.

With respect to the first dimension, aromatase induction (and associated changes in sex steroid ratios) due to atrazine exposure is thought to be a mechanism conserved across the

vertebrate classes. If so, then generalizations from one amphibian clade to others would be appropriate. Similarly, aromatase catalyzation of estrogenesis appears to be highly conserved (Hayes 2005). With respect to the second dimension, can we infer from population level effects of atrazine in one amphibian clade to similar effects in another? In this case, perhaps not, because sex-steroid mediated developmental endpoints may differ among clades (Hayes 2005), and so population level effects will also be likely to differ.

In the atrazine case, we can see that criteria of adequacy drawn from the problem agendas of environmental toxicity and developmental endocrine function are related by the way that they jointly constrain the adequacy of candidate answers to complex questions about atrazine's endocrine disrupting properties and candidate mechanisms.[25]

2.8 The contributions of evolutionary biology

I will now use the criteria of adequacy discussed in section 2.7 to take up a question that was posed in section 2.2: what role does evolutionary biology play in research on the

[25] With respect to these criteria of adequacy, there are not conflicts between the criteria drawn from the problem agenda of developmental endocrine function and the problem agenda of environmental toxicity. However, this is not always the case. There is controversy in atrazine research about the significance of non-monotonic dose responses for atrazine (Rohr and McCoy 2010b; Fagin 2012; Van Der Kraack *et al*. 2014). Such responses have traditionally not counted as demonstrating effects according to criteria of adequacy on dose responses in the environmental toxicity problem agenda, where it has long been thought that "the dose makes the poison" (Vandenberg *et al*. 2012). However, many toxicologists now accept non-monotonic responses as demonstrating effects (Fagin 2012). In contrast, non-monotonic (biphasic) dose responses have long been accepted as evidence of real effects in endocrinology (e.g., Caraty *et al*. 1989). When there are disputes about criteria of adequacy, appeals to more general principles of good science can aid in adjudication. I demonstrate this approach to criteria conflict resolution in Chapter 3.

ecological effects of atrazine as an amphibian endocrine disruptor? First, evolutionary biology can provide population genetic models of amphibian populations, e.g., models of sex ratios in amphibian clades (Wilson and Hardy 2002). These hypothetical populations provide null hypotheses (or baseline characterizations) against which atrazine impacts can be compared.[26] This contribution of evolutionary biology is disclosed by consideration of the *compared to control* criterion of characterizational adequacy.

Second, evolutionary and evolutionary developmental biology provides models of relations among levels of biological organization. Such relations can be understood spatially and temporally both in ontogeny and evolution. Temporal hierarchies in development articulate the relation of, e.g., gene expression to the formation of physiological pathways and morphological structures (2008, 880). In the atrazine case, developmental endocrinology explains how sex steroids at the biochemical level affect the development of sex-specific traits in amphibians at the organismal level. Evolutionary biology contributes here by providing models linking such traits to population level phenomena; population genetic models can articulate relations between organismal traits and population level effects e.g., covariance between abnormal sex ratios as a result of atrazine-induced feminization (aggregated from the sexual character states of individual organisms) and mean fitness in amphibian clades (Hayes 2010; Guiterres and Teem 2006).

[26] See Bausman (2015) for treatment of a distinction between null hypotheses and baseline models.

Finally, evolutionary biology contributes phylogenies of relevant traits, e.g. phylogenies of the CYP19 gene, the sex steroids and their receptors, and phylogenies of certain developmental pathways that are mediated by these steroids. Together, these phylogenies are informative about the degree to which atrazine generalizes as an endocrine disruptor and what its likely effects are across diverse amphibian clades. These phylogenies play an important role in satisfying the *generalization* criterion because such phylogenies can either justify or proscribe inferences from research on one clade to claims about another. Importantly for human health, if aromatase induction and its effects on sex steroids are conserved across vertebrates (Hayes et al. 2011; Hayes 2005; Rohr and McCoy 2010a), this legitimates moves towards seeing amphibian model organisms as sentinel species for threats posed by atrazine exposure to human health.

9. Conclusion

Here I have used Love (2008)'s problem agenda framework to interpret research on the impact of atrazine's endocrine disrupting effects on amphibians as addressing a question located within the problem of assessing the impacts of endocrine disruptors in the environment. This problem is seen as shared by the problem agendas of environmental toxicity and developmental endocrine function. To characterize the epistemic goal of impact assessment central to the environmental problem of endocrine disruptors, I have developed and deployed the concept of criteria of characterizational adequacy, constraints of adequacy on empirically-grounded characterizations and the processes that

generate them. This concept, along with Love (2008)'s concept of criteria of explanatory adequacy, makes clearer the ways in which various disciplines make their contributions to the problem of atrazine toxicity and the question of atrazine's endocrine disrupting effects on amphibians. In particular, we've seen that evolutionary biology contributes by providing models of relevant evolutionary processes and phylogenies (especially of the components of causal chains at multiple levels of biological organization) that inform the propriety of generalizing from findings about one clade to claims about others. Evolutionary biology also contributes by providing models of population-level phenomena that may result from organismal-level atrazine exposure effects.

The forgoing treatment of atrazine research can be seen as a further development of Love (2008)'s and Brigandt (2010)'s response to the challenge issued by Rosenberg (1997) and others. This challenge is for those who participate in the skeptical consensus about the prospects and motivation for Nagelian-type theory reduction to provide alternative accounts of the epistemic relations among scientific disciplines. Love and Brigandt have provided nonreductionist accounts of disciplinary integration centered on solving particular problems and providing particular explanations in evo-devo. Here we've seen how Love's problem agenda framework can be applied to another area of research by expanding this framework to include criteria of characterizational adequacy, criteria constraining what counts as an adequate empirically-grounded characterization given the problems that such characterizations are meant to address. In this way, the forgoing treatment of atrazine research is meant to provide a significant contribution the broader

project of giving plausible nonreductionist problem-centered philosophical accounts of the epistemic relationships among scientific and especially biological disciplines.

Modern research into environmental endocrine disruptors and other complex scientific problems requires multidisciplinary inputs. How it is that these diverse inputs are coordinated and integrated is often unclear. Future enquiry into the criteria of adequacy that unite various scientific problems and problem agendas promises to provide an account of coordination and integration among disciplines and to shed light on the reasons for disagreement when there is a seeming failure to reach consensus. Such an inquiry has the potential for yielding novel interpretations of interdisciplinary disagreement and help to sort disagreement driven by healthy scientific criticism or disciplinary difference from other less reasonable disagreements. By making explicit disciplinary differences with respect to criteria of adequacy associated with various scientific problems, nonreductionist problem-centered epistemologies can help offer resolutions to persistent scientific controversies in a way that honors the contributions of diverse scientific perspectives and practices.

Chapter 3: The Inductive Risk of "Demasculinization"

3.1 Introduction

3.1.1 Endocrine disruption and heteronormativity

There are heated scientific debates about the extent to which pesticides and other chemicals have harmful disruptive effects on the endocrine (hormone) systems of humans and wildlife. The outcomes of these debates have implications for agriculture and industry, regulatory policy, public health, and the environment (Colborn et al 1993; Krimsky 2002; Rohr and McCoy 2010b; Elliott 2011). There are well-publicized and well-motivated concerns that environmental endocrine disruptor debates and policy decisions have been biased by conflict of interest and other inappropriate influences by industry (Aviv 2014; Boone et al 2014). But there is another ethical problem raised by environmental endocrine disruptor debates that has received less attention. The language used by endocrine disruption researchers may be contributing to the reinforcement of scientifically suspect ideas about sex and gender and the maintenance of ethically problematic societal gender norms (DiChiro 2010). Some hypotheses in endocrine disruption research describe the harmful effects of pesticides and other chemicals in gendered terms. When research findings about these hypotheses are presented to the public through popular media and political rhetoric, they are often framed in terms of heteronormative views of sexuality and gender. According to these views, human and animal members of clearly defined binary sexual groups have unique and non-

overlapping sexual morphologies, behaviors, and reproductive roles. Deviations from these heteronorms are often characterized as worrisome or undesirable. A brief (and, compared to other possible examples, benign) quotation from the leading sentence of a *New York Times* article will serve to illustrate:

> Just as frogs' mating season arrives, a study by a Yale professor raises a troubling issue. How many frogs will be clear on their role in the annual springtime ritual? (Barringer 2008)

The study referred to by the *Times* article was eventually published as "Intersex Frogs Concentrated in Urban and Suburban Landscapes." (Skelly et al 2010) The study was focused on variation in frog gonadal morphology, specifically the relative abundance of oocytes (egg cells) in male frog gonads across various landscapes. This study did not investigate the sexual behavior of the frogs. Nonetheless, the lead sentences and conclusion of the *Times* article frame the intervening discussion in terms of the extent to which male frogs in the study deviated from, or were prevented from, engaging in "their role" in reproductive behavior. The *Times* article implies that there is a single reproductive behavior for male frogs, and that this role is determined by morphology. However, Wells (1977, 1978) catalogues several reproductive strategies for males of the frog species in question, *Rana clamitans*. West-Eberhard (1984) and Roughgarden (2009) show how wide variations in sexual behavior found throughout the animal kingdom can contribute to individual fitness effects via "social selection" even when those behaviors do not result in fertilization. Thus, it is unclear the extent to which talk of a unique sexual role for males in natural *R. clamitans* populations is well-supported.

Further, with respect to the significance of the gonadal morphology variations actually measured in the study, Skelley et al (2010) acknowledge that "intersex" gonadal morphology has long been observed in wild frog populations in the absence of significant chemical pollution (Witschi 1921). Barringer's 2008 article does not mention this important fact. This fact does not absolve anthropogenic chemicals of causing increased rates of "intersex" gonadal morphology. However, it does raise the question of the sense in which we should see such morphology as abnormal. Rates of intersex gonadal morphology may be increased by chemical pollution, but intersex gonadal morphology in frogs is only abnormal in the sense that any relatively rare phenotype is. That such morphology does not conform to heteronormative standards does not mean that it is abnormal in any more significant or more worrisome sense.

Similarly problematic media representations of endocrine disruption research are widespread. These representations are often even more explicit than Barringer (2008) in marshaling scientific findings to express anxiety about threats to and deviations from "normal" gender behavior and sexual morphology in both wildlife and humans (Birke 2000; DiChiro 2010). Such language often serves in political discourse as a rhetorical basis for the naturalization of heteronormative social standards. It therefore has the potential to reinforce negative stereotypes of, and exclusionary rhetoric aimed at, people whose sexual morphology or behavior is marked as abnormal according to such standards. Evidence suggests that lesbian, gay, bisexual, and transsexual populations are

at increased risk of discrimination, reduced quality of life, violence, and suicide (Mays and Cochran 2001). "Intersex" individuals have been subjected without consent to harmful "reassignment" surgeries and other abuses (Fausto-Sterling 2000). Kitcher (2002) argues that scientists have an ethical obligation to exercise especial care in accepting hypotheses and pursing lines of research and methodologies that are likely to negatively impact already disadvantaged or oppressed groups. Thus, the use of gendered language in endocrine disruption research is a matter of ethical concern. While one might be tempted to locate the ethical concern exclusively in the practices of the popular media and political actors in representing scientific finding, scientists' choices of language can contribute to the likelihood of problematic political rhetoric and misleading media accounts.

3.1.2 Inductive risk and hypothesis evaluation

Biologists, philosophers, and communications studies scholars have argued that scientific language choices should be evaluated in terms of the potential social impacts of those language choices (Zuk 1993; Nisbet and Mooney 2007; Herbers 2010; Elliott 2009, 2011; Longino 2013). Despite being contested for much of the 20th century, the view that scientists should countenance social, moral and political values in their scientific practices has become widely accepted, in part on the basis of arguments from inductive risk (Brown 2013; Biddle 2013). According to traditional arguments from inductive risk, scientists properly make use of non-epistemic (e.g., social, moral, and political) values in setting criteria to govern scientific reasoning whenever there are significant non-

epistemic consequences of making an error (Rudner 1953; Douglas 2009). On this view, standards of evidence (e.g., confidence levels) are set so as to minimize the risks of type I (false positive) or type II (false negative) errors. Lowering the risk of type I error requires raising the risk of type II error and *vice versa*. Which of the two errors scientists ought to minimize the risk of depends in part upon their respective non-epistemic costs given the non-epistemic values (and the relative weightings thereof) employed by the scientists and scientific communities.[27,28]

Several philosophers have argued that inductive risk can occur at many stages of scientific investigation other than the testing of hypotheses (Douglas 2000, 2009; Wilholt 2009; Biddle 2013).[29] Even in these accounts, however, the inductive risks of earlier stages of scientific investigation are grounded in the impact that choices made at earlier stages have on the likelihood of a hypothesis being erroneously accepted or rejected.[30] Thus, even among philosophers who seek to expand the scope of inductive risk, there remains a tendency to see inductive risk merely in terms of the risks posed by type I and type II errors. This tendency is understandable given the focus that Hempel, who

---

[27] The proper source and weighting of the values employed by the scientists is also contested (Kitcher 2002; Kourany 2010; Brown 2013).

[28] As Douglas (2000) notes, in many circumstances statistical standards are set by convention or choice of statistical software rather than by consideration of reasoned arguments. However, in some cases choices about statistical standards are actually debated by reference to ethical arguments; see Montazerhodjat and Lo (2015).

[29] But see Biddle (2016).

[30] See Elliott and Willmes (2013) for an important discussion of the space of cognitive attitudes that we might take towards hypotheses. For the sake of brevity and simplicity, here I will only reference cognitive attitudes of acceptance and rejection towards hypotheses as a basis for action guidance, e.g., journal publication, advising the media, and policy-making.

popularized the term, placed on type I and type II errors in his most prominent uses of "inductive risk." (1965)[31]

3.1.3 Inductive risk and characterizational choices

But Hempel acknowledged other forms of inductive risk related to choices about scientific terminology, e.g., "the inductive risk of using more than one operational criterion for a given term." (1954, 22) Scientific hypotheses make use of characterizations (concise descriptions) of the phenomena under investigation, and many philosophers accept that there is often a plurality of defensible characterizations of the same or similar phenomena (Dupre 1993; Frigg 2006; Kellert et al 2006; Longino 2013; Biddle 2015; Ludwig 2015). For example, biologists investigating the effects of pesticides on male frog gonadal morphology might describe gonads containing lesions (empty or damaged regions in biological tissues) as "demasculinized" or alternatively simply as "containing lesions." Scientists routinely face such characterizational choices, and some characterizations are better suited than others to fulfill both epistemic and non-epistemic value-based criteria endorsed by individual scientists and scientific communities.[32] This is the case even when the hypotheses constructed on the basis of these characterizational choices are true and correctly accepted as true.

3.1.4 Thesis, argument, and outline

[31] Douglas (2000) and Brigandt (2015) cite Hempel (1965) for the introduction of "inductive risk," though the term appears at least as early as Hempel (1954).

[32] Similarly, Ludwig (2015) argues that choices about scientific ontologies are value-laden.

In this essay, I will argue that characterizational choices pose inductive risks even in the absence of mistakes about the truth of hypotheses.

My central argument is as follows:

1) The concept of induction in the argument from inductive risk is one in which induction is a process for generating action-guidance that is constrained by criteria that are based on favored sets of values.

2) This concept of induction supports a concept of inductive risk as the risk of engaging in a scientific practice that is incongruous with the fulfillment of favored criteria based on favored values.

3) Scientists face choices with respect to how they characterize phenomena.

4) These characterizations are used in the construction of scientific hypotheses.

5) Acceptances of hypotheses containing these characterizations are at risk of failing to fulfill favored criteria based on favored values even if the hypotheses are true.

6) Thus, characterizational choices are locations of inductive risk even in the absence of error with respect to the truth of hypotheses.

In section 3.2, I will review some of the history of the argument from inductive risk and argue that the conception of induction historically at work in this argument demands that we consider kinds of inductive error that can occur even when scientists' judgments

about hypotheses are correct with respect to truth and falsity.[33] In section 3.3, I will introduce characterizational pluralism, the view that there is often a plurality of defensible characterizations of the same or similar scientific phenomena. Section 3.4 will explore gendered language in pesticide research and demonstrate how characterizational choices in inductive processes involve risks even when there is no error with respect to the truth of hypotheses. In section 3.5, I will address an objection to my thesis that such risks are well-described as inductive risks. Section 3.6 will introduce a suggestion for specifying particular inductive risks.

3.2 Turns in the evolution of the problem of inductive risk

3.2.1 Rudner's argument from inductive risk

James Rudner (1953) is responsible for an influential formulation of the argument from inductive risk.[34] Where hypothesis testing is modeled as the comparison of the predictions of a scientific hypothesis to empirical observations, Rudner claims that the acceptance or rejection of hypotheses on the basis of hypothesis testing is an indispensable constituent of "the method of science" (1953, 2).[35] For this reason, Rudner claims that a scientist (in their role as scientist) accepts or rejects hypotheses. Because no scientific hypothesis is ever confirmed with absolute certainty, there exists the possibility

---

[33] For the sake of simplicity and consistency with previous discussions of inductive risk, I will use the language of truth and falsity.

[34] William James is credited with the first formulation of the argument from inductive risk. (James 1896). See also Churchman (1948)

[35] See Jeffrey (1956) for a defense of the view that rather than accept or reject hypotheses, a scientist merely assigns probabilities to hypotheses.

of error. To accept or reject a hypothesis entails a judgment about whether the evidence having bearing on that hypothesis is sufficiently strong to warrant the acceptance or rejection. For Rudner, "our decision regarding the [strength of] evidence and respecting how strong is 'strong enough,' is […] a function of the importance, in the typically ethical sense, of making a mistake in accepting or rejecting a hypothesis." (1953, 2) For Rudner, then, inductive risk is the risk of the erroneous acceptance or rejection of scientific hypotheses in the sense of type I or type II error. Such errors sometimes have ethically significant consequences because some accepted hypotheses guide practical action. Thus, for Rudner, ethical values are an appropriate consideration when setting standards of evidence for the acceptance or rejection of hypotheses.

3.2.2 Douglas and the expanded argument from inductive risk

Douglas (2000, 2009) accepts and provides further argumentative support for Rudner's conclusion. However, Douglas makes an important contribution that broadens the scope of Rudner's conclusion to include other stages of scientific inquiry. Douglas argues that the characterization of data and interpretation of experimental results are also properly locations for considerations of non-epistemic values because choices made at these stages influence whether a hypothesis is accepted or rejected (2000, 2009). On Douglas's view, evaluation of the non-epistemic consequences of error in the acceptance or rejection of hypotheses is appropriate not only to setting standards of evidence, but also to earlier stages of scientific inquiry. Choices made at these stages influence the likelihoods of the acceptance or rejection of particular hypotheses, and thus the likelihoods of the

associated non-epistemic consequences of error when action is taken on the basis of these hypotheses.[36] Note that Rudner and Douglas have implicitly assumed, within the context of the argument from inductive risk, that induction is a process for generating action-guidance that is constrained by criteria that are based on favored sets of values, because it is the action-guiding function of induction that licenses the use of ethical values.

Following Douglas there is increasing philosophical consensus that choices made throughout the various stages of scientific inquiry can affect the outcome of hypothesis testing (e.g., Wilholt 2009), and thereby be locations of inductive risk. But even with the expansion of the argument from inductive risk to cover stages of sciences other than those most closely associated with hypothesis testing,[37] inductive risk is still ultimately viewed in terms the risks of type I or type II error. There is an assumption that the earlier stages of scientific inquiry are locations of inductive risk only by virtue of the effects that they have for hypothesis acceptance or rejection and the associated consequences of making a mistake in judging a hypothesis to be true or false. This is an assumption that should be set aside in discussions of inductive risk.

3.2.3 Hempel, the concept of induction, and a general concept of inductive risk

There is a more general concept of inductive risk as the risk of engaging in a scientific practice that is incongruous with the fulfillment of some favored set of criteria based

[36] Elliott and McKaughan (2009) argue for an even larger expansion of the stages of scientific inquiry that can affect the outcome of hypothesis testing.

[37] See Brigandt (2015) for a critique of attempts to use scientific stage distinctions to circumscribe the permissible roles of values in science.

upon some favored set of values. This conception of inductive risk is consonant with the conception of induction implicit in the history of philosophical discussions of inductive risk. In "Turns in the evolution of the problem of induction" Hempel (1981) reviews the history of philosophical debates about the nature of induction. Of Rudner (1953)'s argument, Hempel says that even if we assume that we are dealing with "pure or basic" science without any ethical implications and thus reject the view that the "scientist *qua* scientist" makes ethical value judgments, it is still reasonable to view the acceptance of a hypothesis into the body of accepted scientific knowledge as itself an action that has consequences (397). However, for Hempel on this model of induction, in the case of "pure or basic" research, these consequences are epistemic rather than ethical and are to be assessed in terms of epistemic criteria based on epistemic values, perhaps, e.g., the set of epistemic values given by Kuhn (1977) including empirical adequacy, simplicity, and explanatory power, among others.[38]

Hempel (1981) was interested in maintaining a distinction between "pure or basic research" and "applied research," and apparently often privileged the former as being the proper subject for inquiry into the nature of induction.[39] However, Hempel closes his 1981 essay by describing an abstract model of scientific inquiry that is consistent with his characterizations of both "pure" and "applied" research. "[S]cientific inquiry aims at

[38] See Rooney (1992), Douglas (1998), and Longino (1995) for critiques of the epistemic/non-epistemic value distinction.
[39] See Kitcher (2002) for a critical account of the distinction between "basic science" and "applied science." See Douglas (2014) for a detailed historical critique.

theories that ever better satisfy certain desiderata, no matter how the latter may be construed in detail[.]" (Hempel 1981, 404)

In the 1965 essay Hempel writes that,

> in a general way, it seems clear that the standards governing the inductive procedures of pure science reflect the objective of obtaining a certain goal, which might be described somewhat vaguely as the attainment of an increasingly reliable, extensive, and theoretically systematized body of information about the world. Note that if we were concerned, instead, to form a system of beliefs or a world view that is emotionally reassuring or esthetically satisfying to us, then it would not be reasonable at all to insist, as science does, on a close accord between the beliefs we accept and our empirical evidence; and the standards of objective testability and confirmation by publicly ascertainable evidence would have to be replaced by acceptance standards of an entirely different kind. The standards of procedure must in each case be formed in consideration of the goals to be attained; their justification must be relative to those goals and must, in this sense, presuppose them. (1965, 93)

So Hempel (in the article in which he made popular the term, "inductive risk") thinks that the standards governing inductive processes must be relative to the goals of the inquiry. But Hempel has here offered a false dilemma with respect to whether epistemic or non-epistemic values are the proper source of these goals and standards. There is currently a philosophical consensus that science and the standards governing scientific practice should be responsive to *both* epistemic and non-epistemic values (Biddle 2013).

Given this general Hempelian model of induction in which inductive procedures are constrained by criteria related to the goals and values associated with inductive projects, philosophers should countenance other ways besides type I and type II error that

hypothesis acceptance can run afoul of favored criteria based on favored sets of epistemic and non-epistemic values, i.e., other forms of inductive risk. I will discuss an example, inductive risk with respect to characterizational choices and ethical values, in section 4. But in order to establish that there are such characterizational choices to be made, first I must introduce characterizational pluralism.

3.3 Characterizational pluralism

3.3.1 A pluralist consensus

Scientific hypotheses make use of characterizations (concise descriptions) of the phenomena under investigation, and increasingly philosophers accept that there is often a plurality of defensible characterizations and classifications of the same phenomena (Dupre 1993; Kitcher 2002; Anderson 2004; Frigg 2006; Kellert et al 2006; Longino 2013; Ludwig 2015; Biddle 2015). Although he does not employ the term, Elliott (2009, 2011) highlights the ways in which characterizational pluralism in environmental research yields language choices that have potential epistemic, social, and environmental impacts that are evaluable in terms of epistemic, social, moral, and political values. For instance, some scientists and policy makers argue that the choice to use the label of "endocrine disruptor" instead of "hormonally active agent" to describe chemicals affecting hormone systems may result in undue worry about relatively harmless classes of chemicals or be prejudicial with respect to open empirical questions about how harmful or disruptive a particular chemical is (Elliott 2009).[40] Similarly Biddle (2015) highlights

[40] See Elliott (2009) for an account of diverse ways that such language choices have social impacts.

the ways that definitional choices for diseases have implications with respect to diagnosis rates. Ludwig (2015) argues that scientists face choices about scientific definitions and that these choices are properly influenced by non-epistemic values. Thus there is a rising consensus among philosophers of science that scientists' choices about characterizations are often not determined solely by the phenomena that they are investigating.

3.3.2 Two forms of characterizational pluralism with respect to scientific language choice

In the case of the use of scientific terminology (*qua* characterization), this characterizational pluralism takes at least two forms: 1) Scientists may use a variety of different terms to describe the same or similar phenomena and 2) Scientists may use multiple definitions of the same term as applied to the same or similar phenomena. For an example of the first sort, and an example that I will analyze in the following section, endocrine disruption researchers might describe male gonadal tissue exhibiting empty or damaged regions as a result of exposure to a pesticide or other chemical merely as "containing lesions," or they might describe such gonadal tissue as "demasculinized." (Hayes et al. 2011) For an example of the second sort, developmental biologists often have different definitional criteria for the proper use of terms like "sterility." On some definitions of "sterility," organisms that produce non-viable embryos are sterile, while on other definitions, "sterile" is only properly applied to organisms that produce no embryos at all (Spike, Bader et al 2008; Spike, Meyer et al 2008).[41]

[41] Huelgas-Morales and Powers (manuscript in preparation) provide a treatment of the characterizational inductive risks of multiple definitions of "sterility" with respect to epistemic values in C. *elegans* germline research.

3.3.3 History and the entrenchment of characterizational choices

Characterizational choices can become entrenched and survive across radical changes in theory. Thus, scientists' characterizational choices are often historically contingent. These choices are also reflective of background societal beliefs and values. To illustrate, according to the historical analysis of Oudshoorn (1994), in the early 20th century societal background beliefs and values influenced the development of two definitional constraints on the binary classification of sex hormones as *sex* hormones, rather than signaling compounds that were not primarily identified with sexual development and function. These constraints were the *criterion of sex-specific origin* and the *criterion of sex-specific function*. These criteria held that in order to be characterized as, e.g., a "female" hormone, the chemical in question needed to originate exclusively in ovaries (as opposed to testes), and needed to exclusively control the development of those morphological features that were taken to be essentially female. By the 1930s, it was clear to many endocrinologists on the basis of available evidence that the chemicals generally did not strictly meet either of these criteria. It was also clear that these chemicals influenced a variety of processes other than just sexual development. There were unsuccessful attempts especially on the part of biochemists to jettison the gendered nomenclature. Plausible factors in the failure of attempts to jettison the sex-based classificatory scheme (despite violations of the criteria upon which the scheme was based) were the

conduciveness of the scheme to cultural beliefs about gender and the related importance of the scheme to the nascent pharmaceutical hormone trade (Oudshoorn 1994).[42]

I now turn to a more recent case exemplifying characterizational pluralism with respect to choices about gendered language and associated criteria (or *criteria of characterizational adequacy*) for the correct use of that language in endocrine disruption research.

3.4 Inductive risk without type I or type II error in endocrine disruption research

3.4.1 Atrazine research and gendered language

Atrazine is a top selling herbicide that is a highly persistent and widely distributed ground and surface water pollutant (Thurman and Cromwell 2000). Recent research has supported the view that atrazine acts as an endocrine (hormone) disruptor in vertebrate organisms. Atrazine exposure is now widely believed to have diverse effects on many different kinds of vertebrate organisms with respect to sexual development (Hayes et al 2011; Rohr and McCoy 2010a).[43] Here I examine some of the non-epistemic considerations salient to acceptance of hypotheses of "demasculinization" of male vertebrate gonads via the endocrine-disrupting effects of atrazine.

Hayes et al (2011) reviews the available evidence on the effects of atrazine with respect to the hypothesis that atrazine "demasculinizes" the gonads of male vertebrates. The authors define "demasculinization" of male gonads as "a decrease in male gonadal

[42] See Richardson (2013) for an analogous case involving the gendered characterizations of chromosomes.
[43] However, many of these conclusions are disputed by studies and researchers funded by atrazine's manufacturer, Syngenta Crop Protection LLC (Solomon et al 2013; Van Der Kraak et al 2014).

characteristics including decreases in testicular size, decreases in Sertoli cell number, decreases in sperm production, and decreases in androgen production." (65) Based upon their use of the term, "demasculinization," the obtainment of any one of the properties given in the definition of the term appears to be sufficient for the application of the "demasculinization" and its cognates. For example, "demasculinized" is properly used to describe the gonads of male vertebrates if empty regions are observed in the gonadal tissue, even if the other properties listed as definitive of "demasculinization" do not obtain.

Significantly, many of the studies that Hayes et al (2011) review do not make use of this gendered characterization of the effects of atrazine on male gonads. Scientists can and have discussed the properties that Hayes et al (2011) list as definitive of "demasculinization" without the use of that or similar terms. For example, McLachlan et al (1975) discuss gonadal lesions in male rats as a result of chemical exposure without the use of the term, "demasculinization" or similar terms.[44] This demonstrates that characterizational pluralism occurs here. Characterizing at least some of the properties that Hayes et al (2011) list as sufficient for the application of "demasculinization" (and its cognates) is a choice that is not necessarily demanded by the nature of the phenomena being described. This choice has consequences that are subject to evaluation in terms of criteria based on both epistemic and non-epistemic values that either are or might plausibly be endorsed by individual atrazine researchers or the atrazine research

[44] See Halina (2015) and Hempel (1958) for treatments of the "theoretician's dilemma" regarding the introduction of theoretical terms. An analysis of the bearing of this dilemma on the case at hand (and more generally the inductive risks of characterizational choices) will have to await future work.

community.

3.4.2 Non-epistemic values and risks associated with language choice in true hypotheses

Consider the hypothesis **D** that the herbicide, atrazine, "demasculinizes" (*sensu* Hayes et al 2011) the gonads of male vertebrates. According to standard versions of the argument from inductive risk, scientists should take seriously the possibility of committing type I or type II errors with respect to **D**. They should evaluate the consequences of these sorts of errors at least in setting standards of evidence (Rudner 1953), and plausibly in other scientific practices (Douglas 2000, 2009; Elliott and McKaughan 2009). If a scientist or scientific community accepts **D** as true when it is in fact false, then that acceptance plausibly increases the likelihood of imposing needless regulatory burdens on the production and application of atrazine. If a scientist or scientific community rejects **D** as false when it is in fact true, then we plausibly increase the likelihood of preventable harms to human health and the environment. According to standard versions of the argument from inductive risk, these are the sorts of outcomes that require scientists and scientific communities to make use of non-epistemic values.

Now suppose scientists accept **D** as true and that **D** is true in the sense that atrazine in fact causes increased rates of male gonadal lesions across a wide variety of vertebrates, one of the properties that Hayes et al (2011) take to be sufficient for the correct application of "demasculinization." There is no type I or type II error in the case so described. But the choice of this characterization of the phenomena of atrazine's production of empty regions in male vertebrate gonads has non-epistemic consequences. Di Chiro (2010)

criticizes what she sees as the heteronormativity implicit in both the scientific language of endocrine disruption research (e.g., "abnormal," "demasculinization" and "feminization") and political rhetoric aimed at limiting the production and distribution of endocrine-disrupting chemicals. People whose sexual morphology and gender behavior are marked as abnormal according to heteronormative standards are represented as harbingers of a toxic environment. Thus, such language potentially serves to reinforce a naturalized account of heteronormativity. Anti-endocrine disruption political rhetoric and sensationalized media accounts marshal scientific findings and the gendered language used by scientists. This rhetoric and these media accounts often capitalize on societal fears of demasculinization, feminization, and gender ambiguity, sometimes by offering lamentable representations of marginalized groups.[45]

On the other hand, suppose that a scientist chooses to eschew the use of gendered terms like "demasculinization" and instead formulate a hypothesis **L** that atrazine causes increased rates of gonadal lesions in male vertebrate gonads. Suppose that **L** is true and scientists accept **L** as true. Again, there is no type I or type II error in the case so described. Nonetheless the choice not to use "demasculinization" in the formulation of the hypothesis plausibly has non-epistemic consequences. As demonstrated by the media analysis of Birke (2000) and DiChiro (2010), there is apparently considerable public anxiety (sometimes taking the form of anti-feminist and anti-LGBTQ rhetoric) about the "feminization" and "demasculinization" of human bodies and culture, as well as the

[45] See Birke (2000) and Di Chiro (2010) for striking examples of these sorts of scientific language, media representations of scientific findings, and political rhetoric. Oudshoorn's (1994) analysis highlights early 20th century attempts to explain "effeminate" men in terms of the action of xenoestrogens.

erosion of conventional gender roles. Given these public sentiments, and assuming the truth of **L**, if scientists were to abandon the use of terms like "demasculinization," they would plausibly thereby forgo some of the potential power of their conclusions in terms of creating effective political rhetoric aimed at limiting the production and application of a harmful chemical. Thus, the use of **L** to the exclusion of **D** may make preventable harms to human health and the environment more likely.

My own moral intuitions about this case incline me to reject the view that scientists ought to use more potentially marginalizing and exclusionary language in order to create effective political rhetoric within the context of a society with heterosexist values. However, the use of such language might well find support within many consequentialist ethical theories. Further, an argument can be made that the use of **D** rather than or in addition to **L** is justified on the basis of non-epistemic values even if we reject the idea of intentionally using language that appeals to heterosexist values. Suppose the truth of the slogan "sex sells," and that people are more interested in scientific findings that they perceive as relevant to their everyday life experience. Suppose further that navigating in, and making sense of, a social world marked by gender diversity is relevant to the lives of most members of the public. Under such plausible suppositions, an argument can be made that the use of **D** is justified because it stands a better chance of being noticed in popular culture, and thereby bringing political attention to the problem of endocrine disrupting chemicals. This example illustrates how different non-epistemic value considerations can favor different language choices.

Suppose that a scientific community accepts a criterion proscribing the use of potentially

exclusionary and marginalizing language based on non-epistemic values of inclusivity and respect for diversity. Given that there is a choice between **D** and **L**, the acceptance of **D** rather than or in addition to **L** runs the risk of failing to fulfill one member of the favored set of criteria that they have endorsed for their research program. On the other hand if a scientific community takes as a criterion minimizing the risk of harms to public health and the environment, then the acceptance of **L** rather than **D** poses other risks within the context of a heterosexist social milieu, or more charitably, a social milieu that cares about gender difference.

It is likely that scientists and scientific communities working on endocrine disruption research value inclusivity and respect for diversity as well as protecting public health and the environment. Thus, this example of non-type I/II inductive risk—and the need to adjudicate between these conflicting values and criteria—is of central and immediate importance. The considerations presented in this subsection should motivate the thought that scientists face trade-offs with respect to fulfilling non-epistemic value-based criteria in their practices. The fulfillment of one set of criteria based on one set of non-epistemic values will often create the risk of the failing to fulfill (or to fulfill as fully) a rival set of criteria based on a rival set of non-epistemic values.

3.4.3 Characterizational choices as locations of inductive risk

In section 3.2.3, I argued that given a general Hempelian model of induction (in which inductive procedures are constrained by criteria related to the goals and values associated with inductive projects), there are other ways besides type I and type II error that

hypothesis acceptance can run afoul of favored criteria based on favored sets of epistemic and non-epistemic values, i.e., other forms of inductive risk. In section 3.4.2, I used the case of **D** and **L** to argue that the choices among hypotheses that reflect alternative characterizations of the effects of atrazine on male vertebrate gonads have potential consequences that are evaluable in terms of non-epistemic values. Assuming that endocrine disruption communities endorse values including inclusivity and respect for diversity, as well as protecting public health and the environment, and endorse criteria for scientific practice based on these values, choices among **D** and **L** constitute inductive risks with respect to this set of values and criteria.

3.5 Biddle's schema objection

Here I will address an objection to my claim that the kinds of characterizational choices exemplified in the preceding sections constitute locations of inductive risk. Biddle (2015) claims that we should only label as inductive risks those risks that fit within what he takes to be the schema of the traditional argument from inductive risk. Biddle claims that the focus of the argument from inductive risk is on the decision of how much evidence is sufficient to accept or reject a hypothesis given the possibility of being mistaken about the truth of that hypothesis.[46] He argues that in order to preserve conceptual clarity, philosophers should reserve the term "inductive risk" for those

[46] This claim is true of Rudner's account, which does indeed focus on the amount of evidence. It is less obviously true of Hempel's account, which focuses on "how strong the evidential support has to be" in order to accept a hypothesis. It is not clear that Hempel's account is merely about the amount of evidence, since the strength metaphor might also imply considerations other than the amount of evidence, e.g., the quality of the evidence.

arguments that deal with the roles of values in determining how much evidence is sufficient to accept a hypothesis. For this reason, he rejects, for example Douglas (2009)'s claim that there is inductive risk in choices about the characterization of data[47] and Wilholt (2009)'s claim that there is inductive risk in the choice of model organism.

Both Douglas and Wilholt emphasize the implications of scientific decisions in terms of erroneously accepting or rejecting hypotheses in their extensions of inductive risk. My thesis that characterizational choices are locations of inductive risk does not appeal to mistakes about the truth of hypotheses. Thus, my arguments are even further removed than Douglas's and Wilholt's from what Biddle takes to be the traditional argument from inductive risk. This implies that my arguments and thesis are an even greater threat to the conceptual clarity of inductive risk than are Douglas's and Wilholt's.

The first thing to note is that it is not clear that *conceptual* clarity is preserved by restricting the use of the term "inductive risk" to refer to decisions about standards of evidence. As I have argued in sections 3.2.2 and 3.2.3, the concept of induction in the argument from inductive risk is one of an action-guiding process that is constrained by criteria that are based upon values or goals associated with inductive projects. Given this concept of induction, and the fact that inductive projects are constrained by criteria based upon diverse epistemic and non-epistemic values, the concept of inductive risk should

---

[47] Douglas's concerns about characterization of data differ from my concerns about characterizational choices in the following respect: Douglas has focused on choices about how to code ambiguous data. For example, there is a choice to be made about whether to say that a liver sample is cancerous or non-cancerous when experts disagree about whether the sample is cancerous or non-cancerous (2000, 2009). In Douglas's primary example involving liver samples, the terminology is fixed, and the choice is about which of two mutually exclusive terminological categories to subsume the data under. I am focusing here on cases in which there is a choice to be made about the kinds of terminology used to describe a phenomenon. These choices need not involve any ambiguities or expert disagreement about what the data show.

include the various ways that decisions made during the course of inductive projects can fail to fulfill the criteria that govern the projects. The concept of inductive risk seems to be obfuscated rather than clarified by restricting its scope to concerns related to the amount of evidence necessary to accept a hypothesis. Such a restriction implies an implausibly simple model of induction as mere evidence counting. As I argued in section 3.2.3, Hempel's own account of induction was more complex.

As we saw in section 3.3.3, with the case of gendered hormone classification, characterizational choices can become firmly entrenched in science for historical reasons even when subsequent inquiry shows that there are reasons for altering the entrenched characterization. The same holds true for philosophy. Hempel used the term "inductive risk" to refer to terminological choices (1954) as well as rules of acceptance (1965). The 1965 article and the use of "inductive risk" found there have been very influential and the term "inductive risk" has become most associated with decisions about evidence-sufficiency rules for hypothesis acceptance. However, this historical consideration does not imply that the concept of inductive risk is fundamentally about the sufficiency of evidence.

It is plausible, however, that expanding the scope of inductive risk in the ways advocated by Douglas, Wilholt, and myself do represent a threat to the *terminological* clarity of and thus usefulness of the term, "inductive risk." I address this concern in the concluding section.

3.6 Conclusion

Douglas (2009), Wilholt (2009), and I have argued that there are inductive risks with respect to decisions other than those concerning the amount of evidence necessary to accept a hypothesis. These decisions include decisions about the characterization of data, interpretation of evidence, choice of model organism, and terminology describing the phenomena being investigated. That these and other decisions are locations of inductive risk does suggest the need for more fine-grained descriptions of inductive risk. Declarations that some scientific decision is a location of inductive risk should be met with the question, "Inductive risk with respect to what?"

Given the diversity of inductive practices, goals, and criteria, the construction of a general taxonomy of inductive risks is beyond the scope of this paper. However, we can think of inductive risk as a relational concept describing relationships between values, value-based criteria, and decisions about scientific practices. Answers to the question "inductive risk with respect to what?" should therefore specify how particular decisions in scientific practices pose a risk of failing to fulfill values or value-based criteria governing the scientific practice in question.

To illustrate, in section 3.4.1 I established that endocrine disruption researchers face choices about whether to use gendered language to describe some of the effects of endocrine disrupting chemicals. In section 3.4.2, I showed how choices to either use or eschew the use of gendered language pose risks with respect to criteria like avoiding the promotion of heterosexist gender norms and protecting the environment against harmful chemicals.

The risks described in section 3.4.2 are inductive risks, in the sense that these choices about characterizations pose risks with respect to the values and value-based criteria that either actually are or plausibly ought to govern the inductive processes of endocrine disruption research. These characterizational choices are in this sense, inductive risks, even if there are no mistakes about the truth of the hypotheses that contain the terms resulting from these characterizational choices.

Chapter 4: Global Demarcations, Local Criteria, and Evidence of Bias

4.1 Introduction

Philosophical discussions of ideologically biased science have focused on the project of providing "some general criteria… for articulating conditions under which scientific research may be inappropriately impacted by political ideals or economic interests." (Steel and Powys Whyte 2012) Call this sort of approach to questions about values in science a global demarcation project. Global demarcation projects generally have two aims. The first is the theoretical aim of providing criteria for distinguishing good science from problematically biased science and can be understood as addressing a version of the demarcation problem: What features do all instances of epistemically good science and no instances of epistemically bad science have in common? The second is the practical aim of science criticism, saying whether or not particular instances of scientific reasoning and practice should be discounted as being compromised by illicit influences of values (e.g., financial conflicts of interest). Prominent strategies for globally separating good science from problematically biased science appeal to distinctions between direct and indirect roles for values in science (Douglas 2009), distinctions between social and cognitive values (Lacey 2005), and distinctions between epistemic and non-epistemic values (Steel and Powys Whyte 2012). Global demarcation projects rest on an implicit

assumption that there are generally applicable good-science-making and bad-science-making features to be discovered, articulated, and used to make decisions about when to discount science as problematically biased. In this paper, I will argue that global demarcation projects as currently undertaken are unlikely to meet either of their aims and suggest an alternative approach. The alternative approach reinterprets global demarcation projects as providing prima facie principles of good science rather than exceptionless globally applicable principles. Prima facie principles resulting from modest demarcation projects are to be integrated with appeals to local criteria of adequacy for scientific practices, and principles of inference for detecting the illicit influence of values in science. All three of these components (prima facie principles of good science, local criteria of adequacy, and principles for inferring bias) are necessary to address pressing practical problems related to the biasing effects of financial conflicts of interest in science.

Global demarcation projects as currently undertaken suffer from three limitations. First, recent work on the nature of concepts has called into question the assumption that concepts like SCIENTIFIC METHOD are amenable to being well-described by a small set of globally applicable and informative definitional properties (Vickers and Taylor 2015).[48] The history of debates about the nature of SCIENCE- and all other complex concepts- is a history of failed attempts at providing such properties (Hansson, 2015). This suggests that any proposed global demarcation principles will either be subject to

[48] I follow the conventions of denoting concepts by using capital letters and denoting terms referring to concepts by enclosing them in single quotation marks.

counterexample or too vague to be informative about whether particular cases of scientific practice are inappropriately biased.

Second, global demarcation projects do not give sufficient attention to the diversity of criteria employed by scientists to identify problematic science within their areas of expertise. Scientific communities have *criteria of adequacy* that constrain what counts as adequate scientific descriptions and adequate scientific explanations (Brigandt 2015; Powers 2014). Such criteria are deployed in actual scientific debates in order to identify science that does not meet the standards of the scientific community involved in the debate.

Third, global demarcation projects focus on claims about the good-making and bad-making features of scientific reasoning with respect to values. They have not focused on providing principles for inferring when particular cases of problematic science are likely to be due to the pernicious influence of, e.g. financial conflicts of interest, rather than honest mistakes, legitimate differences of opinion, inadvertent methodological deficiencies, *etc*. Without these principles of inference, global demarcation approaches can at best only give conditions under which science is problematically value-laden. They cannot tell us whether we have evidence that the illicit influence of values is the likely cause of methodologically problematic science.

In order to overcome these limitations, global demarcation projects must be reinterpreted and augmented with two kinds of complementary projects. Call the first local criteria projects. Such projects aim to uncover evidence of problematic reasoning in particular scientific debates without assuming that there are properties had globally by all instances of good science. Local criteria projects focus on criteria of adequacy deployed in particular scientific discourses rather than more abstract global standards for the demarcation of good science and bad science.

The second kind of complementary project is focused on specifying conditions under which it is reasonable to infer that certain kinds of values are playing an inappropriate role in scientific reasoning. Call these evidence of bias projects. I focus on financial conflicts of interest and propose that the strength of evidence of problematic bias via financial conflicts of interest increases as more members of a set of conditions (given in section 4.4) are met.

In section 4.2, I give an overview of reasons for skepticism about providing global criteria for differentiating good science from bad science. I show how two prominent global demarcation projects, the direct roles approach (Douglas 2009) and an approach that favors epistemic over non-epistemic values (Steel and Powys Whyte 2012) are respectively subject to counterexample and too vague to fulfill the practical aim of science criticism. In section 4.3, I introduce local criteria projects, and show how appeals to local criteria can identify problematic methodological practices using a case drawn

from endocrine disruption research. Section 4 proposes a set of principles for inferring bias by financial conflicts of interests.

In section 4.5, I close by offering a proposal for integrating modest demarcation projects, local criteria projects, and evidence of bias projects, and highlight some of the motivations for and challenges facing each kind of project. I focus on the complementary roles that the three kinds of projects play. Local criteria rather than demarcation principles will generally be the most relevant standards for recognizing methodologically problematic science. However, a set of modest demarcation principles (when conceived of as *prima facie* principles rather than universally applicable exceptionless principles) will serve to help adjudicate conflicts with regard to contested or controversial local criteria, and help to identify problematic research programs whose criteria of adequacy conflict with broadly recognized principles of good science. Evidence of bias principles will generally be most relevant to the goal of determining when methodological problems in science are likely to be due to the illicit influence of values rather than honest mistakes, reasonable methodological disagreements, ignorance, *etc*.

4.2 Concepts and Global Demarcation Projects

4.2.1 The death of the definitional view of concepts and the rise of pluralism.

Traditionally, concepts have been thought of as definable by sets of properties that are individually necessary and jointly sufficient to capture the meaning and extension of the concept in question. Suppose the definition of 'mud' (referring to the concept, MUD) is earth mixed with water. This definition aims to capture the meaning of 'mud' and to set conditions for what properly counts as mud. Call this the definitional view of concepts, where the presence of counterexamples is thought of as evidence that a definition is faulty. Counterexamples are cases that either 1) meet the definition, but do not intuitively count as instantiations of the concept, or 2) do not meet the definition, but intuitively count as instantiations of the concept. Call this the extensional adequacy criterion on definitions (Taylor and Vickers 2015). Sprinkle a pinch of dirt into a large glass of water, and you have defeated the definition of 'mud' as earth mixed with water. The counterexample shows that this definition of 'mud' does not meet the extensional adequacy criterion because a relatively large amount of water mixed with a relatively small amount of earth does not intuitively seem like mud.

Taylor and Vickers (2015) give a set of 20 examples of concepts drawn from diverse domains of inquiry that resist exceptionless definition. Philosophical concepts that resist definition include LOGICAL CONSEQUENCE, ART, CONCEPT, and NATURAL KIND among others. Scientific concepts that resist exceptionless definition include ACID, MEMORY, HEALTH, and SPECIES among others (Taylor and Vickers 2015). Based upon the history of failed attempts at exceptionless definitions for complex

concepts, it appears that counterexamples are similar in the relevant respect to death and taxes.

A plausible reason for the resistance of these concepts to exceptionless definitions is that the definitional view of concepts is probably mistaken. This at least is a near consensus view in the philosophy of concepts (Fodor 1998; Machery 2009) and the psychology of concepts (Margolis and Laurence 1999). And certainly we have strong inductive evidence that concepts resist being defined in ways that meet the extensional adequacy criterion. In the long history of philosophy, there are no uncontroversial examples of non-stipulative definitions of complex concepts that meet this criterion (Fodor 1998).

Instead, we see a move in diverse domains of philosophy and science towards a kind pluralism (Taylor and Vickers 2015). Jettisoning the global extensional adequacy criterion, we see that different definitions of concepts are useful in different contexts and for different purposes. Consider Wilson (2006)'s treatment of HARDNESS in material science. Wilson notes that in evaluating the hardness of objects in everyday life and in scientific and engineering applications, we typically use different testing methods depending upon the object and the purpose of the evaluation. The different tests of hardness seem to travel with associated criteria of hardness and these tests and criteria can be more or less adequate for particular purposes. We evaluate the hardness of wood differently than we evaluate the hardness of flowerpots, and we evaluate the hardness of wood differently when we are searching for a scratch-resistant veneer instead of a sturdy

structural timber. Many test-specific operational definitions of hardness are incompatible insofar as they generate different ordinal rankings of the hardness of materials. E.g., a scratch test might indicate that material A is harder than material B while a test of resistance to indentation might indicate that material B is harder than material A. Once we contextualize the meaning of 'hardness' by specifying, e.g., the material, the sorts of tests, and the applications to which the material will be employed in light of testing, it seems likely enough that we could specify a limited though useful definition for 'hardness,' but only under these circumstances. In short, definitional criteria for 'hardness' in materials science will be localized to particular testing procedures, test goals, and materials.

This sort of pluralism is now either a rising or consensus view about a wide range of philosophical and scientific concepts (Taylor and Vickers 2015), including SCIENTIFIC METHOD (Sankey 2010).

4.2.2 Related difficulties for global demarcation projects

There is reason to think that philosophers of science engaged in global demarcation projects have not properly countenanced the death of the definitional view of concepts. Global demarcation projects still aim at accounts of the proper roles of values in science that are applicable to all of science in context independent ways. I illustrate with the examples of Douglas's direct roles approach (2009) and the epistemic value-focused

approach of Steel and Powys Whyte (2012). These approaches issue demarcation principles that are respectively subject to counterexample and too vague to be informative about distinguishing cases of problematically biased science.

In an influential treatment of the proper role of values in science, Douglas (2009) attempts to demarcate "unacceptable politicized science" from science that she takes to be unproblematically value-laden. The former occurs when "values are allowed to direct the empirical claims made by scientists." Douglas's view depends on a distinction between *direct* and *indirect* roles. Values play a direct role when they "act as reasons in themselves to accept a claim, providing a direct motivation for the adoption of a theory." (p. 96) In contrast, values play a proper indirect role when, for example, they "act to weigh the importance of the uncertainty of the claim, helping to decide what should count as sufficient evidence for the claim." (p.96)

It is unclear the extent to which Douglas takes herself to be providing a monistic definition of some concept like POLITICIZED SCIENCE. Douglas's pluralism about the concept of OBJECTIVITY suggests that she does not (or perhaps should not) (2009, Chapter 6). However, Douglas is explicit that her view of the distinction between unproblematically value-laden science and politicized science "should cover *all* of science." (2009, 113-114) This suggests some degree of monism about POLITICIZED SCIENCE. It also suggests that failures with respect the criterion of extensional

adequacy, as evidenced by counterexamples, count against her account of POLITICIZED SCIENCE by her own lights. I will now sketch such a counterexample.

In 2012, agency scientists at the USEPA concluded that the herbicide atrazine does not affect the sexual development of amphibians in the wild (USEPA 2012, 62). This decision had important implications for atrazine's manufacturer because a contrary finding would have been likely to place additional regulatory burdens on the application of atrazine and thus negatively impacted atrazine sales. However, the agency appears to have reached their conclusion on the basis of a single study, Kloas *et al*. (2009). This single study was funded by atrazine's manufacturer, Syngenta. There were 74 other relevant studies available, many of which did find a link between atrazine and abnormal amphibian development. The reason that these 74 studies were excluded from consideration is that they did not meet the methodological standards of the agency for evaluating pesticide risk (Boone et al. 2014). According to a paper coauthored by scientists who served on an EPA Scientific Advisory Panel for atrazine risk-assessment,

> The USEPA works with industry to establish the methodology and experimental design for studies. The complexity and logistics of these designs can make them prohibitively expensive for researchers outside of industry, often leaving industry as the only entity that can afford to conduct the research to the USEPA's specifications or that is knowledgeable of the requirements. Therefore, all or most of the data used in risk assessments may come from industry-supplied research, despite clear COIs [conflicts of interest]. (Boone *et al*. 2014)

Assuming that the assessment of Boone *et al*. is correct, this seems intuitively to be a case of politicized science, especially given widespread evidence that funding is a predictor of

study outcome in toxicology in general, and atrazine research in particular (Hayes 2004; Elliott and Resnik 2014; Bero et al. 2015). EPA scientists were constrained by the requirement that they only consider scientific findings that meet agency standards. Industrial interests were allowed to influence these standards such that only industry-funded science was capable of meeting them, thus excluding science that was not in the control of industry. [49] Yet setting standards for what counts as good evidence seems to be a case of values playing only an indirect role in the scientific reasoning, not a direct one. From the perspective of the agency scientists, the values of industry did not "act as reasons in themselves to accept a claim." (Douglas 2009) But the values of industry did act "to decide what should count as sufficient evidence." (Douglas 2009) And this decision led to the exclusion of studies that were more likely to find effects of atrazine. This case thus serves as a counterexample to Douglas's account of POLITICIZED SCIENCE. Given the death of the definitional view of concepts, such counterexamples are to be expected.[50]

In contrast, Steel and Powys Whyte (2012) advance an alternative "general values in science standard" (hereafter, GVSS) that "nonepistemic values should not conflict with epistemic values in the design, interpretation, or dissemination of scientific research that is practically feasible and ethically permissible." (164) On their account, epistemic values

[49] My argument does not depend on the claim that the industry-influenced EPA standards are bad standards. These standards include "the use of glass containers, loading densities lower than 1 tadpole per liter, measured pesticide or ammonia concentrations in treatments and controls throughout the study, [and] the use of a flow-through design apparatus." Boone *et al*. (2014)
Even if meeting the standards would be epistemically desirable, one can arguably still draw well-supported conclusions from studies that do not meet them.

[50] Steel and Powys Whyte (2012) and Brigandt (2015) offer other counterexamples to Douglas's direct role demarcation criterion.

are values that are "truth-promoting" and nonepistemic values are those that are either neutral with respect to or impede the attainment of truth.

The first thing to notice about GVSS is that it is not at all clear under what circumstances non-epistemic values should not conflict with epistemic values. The difficulty lies in the qualification that non-epistemic values should not conflict with epistemic values in "scientific research that is practically feasible and ethically permissible." Since it would seem that non-epistemic values are the standards according to which we evaluate whether scientific research is practically feasible and ethically permissible, we can restate GVSS as "nonepistemic values should not conflict with epistemic values in the design, interpretation, or dissemination of scientific research that is acceptable in light of non-epistemic values." This does not seem very informative. [51] Because the principle as restated is permissive, it is difficult to think of clear counterexamples. But immunity from counterexample is bought at a cost. While GVSS serves as a potential explanation of why biased science is biased, because it is so permissive, it does a poor job of picking out cases of biased science.

Consider the following case developed by Steel and Powys Whyte (2012).

[51] One might think that GVSS can be improved upon by assuming a multi-stage process. In the first stage there is a decision about whether the research is practically feasible and morally permissible, and where nonepistemic values can take priority over epistemic values. In the second stage, epistemic values are to have priority with respect to decisions about the design, interpretation, or dissemination of the research. While this strategy would more clearly delineate the respective roles of epistemic and nonepistemic values, it seems difficult to reconcile with scientific practice. In the course of research, new practical and ethical challenges are liable to emerge. So researchers might be frequently toggling back and forth between the two stages in ways that would make them difficult to disentangle. See Brigandt (2015) for related arguments against Douglas's normative use of stages. Thanks to Naomi Scheman for suggesting this possibility.

> a pharmaceutical corporation… thinks that it is much worse to mistakenly infer that a drug is ineffective than to erroneously conclude that a drug is effective. The corporation may make this value judgment because it is concerned to generate profit but also because its members sincerely, though perhaps self-servingly, believe that the human benefits of new pharmaceuticals generally outweigh the risks. So, the corporation demands a much higher standard of evidence for accepting, and hence publishing, studies with negative results. (171)

Steel and Powys Whyte claim that setting relatively high standards of evidence for negative results conflicts with epistemic values that demand "severe" tests of hypotheses and so conflicts with the GVSS. But this conclusion is far from clear. If the corporate members sincerely believe that the benefits of new pharmaceuticals generally outweigh the risks, then given this belief it would be arguably unethical (on consequentialist grounds) to accept lower standards for accepting negative results. Thus, there is a case to be made that the activities of the corporation would not violate GVSS because accepting negative results based on lower standards would be ethically impermissible. Due to its permissiveness, GVSS cannot say whether this case is one in which the use of values are illicit. GVSS still serves as potential theoretical account of constraints on the appropriate roles of values in science, but because of its permissiveness, it appears to be ill-suited to the practical goal of distinguishing particular instances of licit and illicit influences of values in science, even when, as stipulated in the case, we know the motivations of the parties.

But there is a deeper problem with GVSS that threatens its suitability with respect to the providing a theoretical account of constraints on the appropriate roles of values in

science. There is reason to suspect that there are few values save perhaps, trivially, internal consistency that are truth-promoting in a context independent manner. Take, for example, the central epistemic value of empirical adequacy. Empirical adequacy is a fit between the predictions of a model and observed data. But when a false (but believed to be potentially true) model has predictions that match observed data, empirical adequacy does not promote true beliefs about that model. Consider the case of a false model whose predictions match with a biased data set. In the atrazine case above, from the perspective of the EPA agency scientists, the null hypothesis that atrazine had no effect on amphibians was a better fit to the observed data generated in Kloas *et al*. (2009). But given empirical evidence about the effects of funding on toxicology research, there is at least some reason to question the extent to which the data provided in Kloas *et al*. are representative of the effects of atrazine on amphibian gonads (Hayes 2004; Bero *et al*. 2015). *If* the data *are* problematic, then empirical adequacy is not truth-promoting in this context. Given the death of the definitional view of concepts, we should not expect to find values that are truth-promoting in non-trivial and context-independent ways.

So it is with Steel and Powys Whyte's primary example of an epistemic value, *severity,* which requires that good evidence in favor of a hypothesis must result from a test procedure with a high probability of uncovering the falsity of the hypothesis if the hypothesis is false. Suppose that some chemical causes cancer in humans, and that the hypothesis that this chemical causes cancer is supported by evidence from 1) an epidemiological study providing a correlation between human populations exposed to the

chemical and the presence of cancer and 2) a study providing the mechanism by which the chemical leads to cancer formation in rats. Now suppose that the manufacturer of the chemical objects to the acceptance of the hypothesis that the chemical causes cancer in humans by invoking severity. The epidemiological test is not properly severe because factors other than the chemical might be causing the cancer in the sampled population, and the rat study is not properly severe because the mechanisms that give rise to cancer formation in rats may not be present in humans. Since none of the tests purporting to provide evidence for the hypothesis are properly severe, says the manufacturer, there is no good evidence in favor of the hypothesis. In this case, demanding severe tests was not truth-promoting, since, as stipulated, the hypothesis was true, but demands of severity were marshaled to promote skepticism about that true hypothesis. This kind of promotion of skepticism on grounds of severity is a common tactic of industry to avoid regulatory burdens. This tactic has been deployed in cases involving tobacco, acid-rain, ozone depletion, and climate change (Oreskes and Conway 2011). While the prima facie principle that tests of a hypothesis should be severe is plausible, the grounding of that plausibility cannot be the propensity of severity to promote truth in context independent ways. The truth-promoting capacities of the severity principle are not independent from how the principle is used.

But suppose that we grant that severity is always truth-promoting. Is the severity principle well equipped to distinguish problematic scientific practices that might indicate the illicit influence of values? Consider Steel and Powys Whyte's diagnosis that a

prominent study in environmental justice research violated the severity principle, the principle that states that "data x_o do not provide good evidence for hypothesis *H* if x_o result from a test procedure with a very low probability or capacity of having uncovered the falsity of *H* (even if *H* is incorrect)." (2012, 168) Anderton *et al*. (1994) was an industry-funded study that, in contrast to previous studies found no significant relationship between the racial composition of neighborhoods and the location of toxic waste facilities. Drawing on Mohai (1994), Steel and Powys Whyte claim that Anderton *et al*. did not provide a properly severe test of the hypothesis that toxic waste site location is correlated with race because of problematic methodological choices about the control group and units of analysis in the study.

Environmental justice researchers face choices about whether to use single census tracts or aggregated census tracts as the units to be compared with respect to correlations between racial demographics and toxic waste site location. Anderton *et al*. claimed that there was no reason to prefer aggregated census tracts, and thus chose to perform their analysis based on single census tracts. Citing Mohai (1994), Steel and Powys Whyte write that Anderton *et al*. should

> have also chosen larger units of analysis than [single] census tracts. Units of analysis in studies that test hypotheses concerning environmental injustices relating to the locations of locally undesirable land uses (LULUs) should approximate the area affected by the site. Mohai pointed out that several studies relevant to this topic published prior to [Anderton *et al*.] (e.g., Nelson *et al*. 1992) had found that residential property values were depressed within 2.0 to 2.5 miles of the LULU (1994, 629–32).

With respect to the control group in Anderton *et al*., the study excluded rural areas and urban areas that did not contain at least one toxic waste site. Steel and Powys Whyte (2012, 176) write that

> To properly test such a statistical hypothesis one needs a sample that includes a representative collection of areas both with and without TSDFs [toxic waste sites] so that it is possible to check whether those with TSDFs have a larger proportion of minority residents than those without. Excluding metropolitan areas without TSDFs makes the study much less likely to find any such correlation[.]

Steel and Powys Whyte conclude that the severity principle was violated in Anderton *et al*. The choice of units of analysis and the choice of control group led to a test that was not properly severe to provide good evidence.

Notice that in the environmental justice case described above, the severity principle does not, strictly speaking, apply. In their published work Anderton *et al*. were not explicitly providing evidence *for* a hypothesis. Rather, the conclusion of their study was that there was not sufficient evidence of a correlation between race and toxic waste site location. Anderton *et al*. merely claimed that they were unable to reject a null hypothesis of no correlation based upon the data generated by their study. So the severity principle as written cannot explain the putatively problematic nature of the study, since the severity principle governs the conditions in which we have evidence *for* a hypothesis.

But suppose that Anderton *et al*. or their sponsors took the study to show that there was positive evidence of no-correlation rather than merely the failure to reject no-correlation.

Even under this supposition, it is not clear that there was a severity principle violation. In order to show that there was a severity principle violation, Steel and Powys Whyte need to show that the test procedure in Anderton *et al*. has "a very low probability or capacity" of finding a correlation between race and toxic waste sites even if there is actually such a correlation. But Steel and Powys Whyte have only shown that, given certain background assumptions, methodological choices about control groups and units of analysis in Anderton *et al*. have made it *less likely* to find a correlation. But this does not yet show a severity principle violation. In order to establish that, we would need some criterion for what counts as "a very low probability or capacity" of finding a correlation between race and toxic waste sites even if there is actually such a correlation.

A natural reply would say that "a very low probability or capacity" of finding a correlation between race and toxic waste sites obtains when the methodological choices are such that the effect disappears. We could then reformulate the severity principle to state that "data x_o do not provide good evidence for hypothesis *H* if x_o result from a test procedure that does not actually uncover the falsity of *H* if *H* is incorrect." But in order to establish a violation of the principle as restated, we would need to know that the hypothesis in question is incorrect. Knowing this would seem to obviate the need for the severity principle in the case. The goal of the severity principle is to provide a criterion of good science, and reformulating the severity principle in this way reduces us to saying that good science is science that gets it right.

Another option would be to say that the severity principle is violated when the test procedure has a merely reduced or sub-optimal capacity for uncovering false hypotheses. But this is surely implausible, since routine and defensible choices about, e.g., confidence levels, can have the effect of reducing the likelihood of uncovering a false hypothesis.

Perhaps there are other options, but given the death of the definitional view of concepts and the rise of pluralism, we should probably not expect to find a criterion of "low probability or capacity" that is universally applicable across diverse scientific contexts. But without such a criterion, the severity principle as given by Steel and Powys Whyte (2012) cannot account for the problematic nature of methodological choices in Anderton *et al*., since we cannot establish severity principle violations.

Fortunately, we do not need the severity principle to account for the methodological shortcomings of Anderton *et al*. There are well-accepted criteria governing statistical hypothesis testing that appear to have been violated in Anderton *et al*. In statistical hypothesis tests of correlation, samples drawn from both the control population and the test population should be representative of that population, or at least as representative as possible given the available methodologies. Given a choice between two or more approximately equally practically-feasible methodologies, researchers testing a statistical hypothesis should chose the methodology that produces the most representative samples. In Anderton *et al*., samples of the control population were not representative because they excluded non-urban areas and urban areas without a toxic waste site. They could have

easily included such areas. Anderton *et al*.'s samples of the test population were not representative because they excluded impacted areas not contained within the single census tracts. Previous work had indicated that a 2.5 mile radius is a robust guideline for the impact zone of toxic waste facilities using depressed home values as an indicator (Nelson *et al*. 1992). Anderton *et al*. could have easily aggregated census tracts so as to better approximate the area impacted by toxic waste sites. So we do not need the severity principle or a criterion of "low capacity" to establish and explain the methodological short-comings of the study. We merely need to appeal to well-established criteria governing tests of statistical hypotheses, well-confirmed background knowledge accepted in the domain of inquiry, and the array of methodological choices available to the researchers. In the following section, I elaborate on criteria of adequacy and provide another example of their usefulness in identifying problematic scientific reasoning.

4.3 Local Criteria Projects

4.3.1 Values and criteria of adequacy in science

Brigandt (2015) emphasizes the roles of values in establishing *criteria of adequacy* for scientific theories. Criteria of adequacy for scientific theories are criteria endorsed by scientists and scientific communities on the basis of both epistemic and non-epistemic values that specify what counts as an adequate theory relative to both the scientific community in question and the scientific problems that the theory is meant to address.

For example, empirical adequacy is a widely shared epistemic value because most theories have as a goal a fit between the predictions of models or theories and experimental results. According to the value of empirical adequacy, there ought to be a fit between the predictions of a theory and experimental data. This value generates criteria of adequacy (e.g., measures of statistical significance), the fulfillment of which ensures that the value of empirical adequacy exerts proper influence on the relevant scientific practices and conclusions.

In the following section I will illustrate the role of these criteria in diagnosing problematic science using the example of dose-response graphs in toxicology. These graphs are outcomes of experiments that have the proximal goal of giving a characterization of organismal responses to toxins. Dose-response graphs must meet "concentration-relative" and "compared-to-control" criteria (Powers 2014). These are criteria of adequacy that constrain acceptable characterizations of the response of a biological entity to a putative toxin.

4.3.2 Criteria of adequacy as a tool for diagnosing methodologically problematic science

Appeals to local criteria of adequacy can identify problematic science. Consider conflicting interpretations of an industry-funded study (Coady *et al*. 2004), on the effect of the herbicide, atrazine, on metamorphosis, growth, and gonadal morphology in the green frog, *Rana clamitans*. Two review articles (Van Der Kraak *et al*. 2014; Solomon *et al*. 2008) funded by atrazine's manufacturer, Syngenta Crop Protection LLC (hereafter, Syngenta), have described Coady *et al*. as showing atrazine having no effect on several

developmental endpoints. In contrast, Rohr and McCoy (2010b) claim that because the controls in Coady *et al*. were contaminated with atrazine, such claims of no effect are specious. Many of the coauthors in Van Der Kraak *et al*. and Solomon *et al*. have endorsed in other contexts a widely-shared compared-to-control criterion that was not met in Coady *et al*., and thus appear to be selectively applying this criterion.

Coady *et al*. exposed treatment groups of larval *R. clamitans* to nominal concentrations of atrazine of 0 (control), 10, and 25 μg/L. They measured the treatment groups with respect to age at metamorphosis, among other developmental measures. In order to verify that actual levels of atrazine exposure matched nominal levels, tank water from the treatment groups was analyzed using immunosorbent assay and gas chromatography-mass spectroscopy. These analyses showed that the atrazine concentration in the control group was between 0.07 and 0.25 μg/L (Coady *et al*. 2004). Measurements of the treatment groups showed statistically significant differences ($p<0.05$) between the nominal 10 μg/L atrazine treatment group and the nominal atrazine control group with respect to age at metamorphosis (Coady *et al*. 2004). See Figure 1.

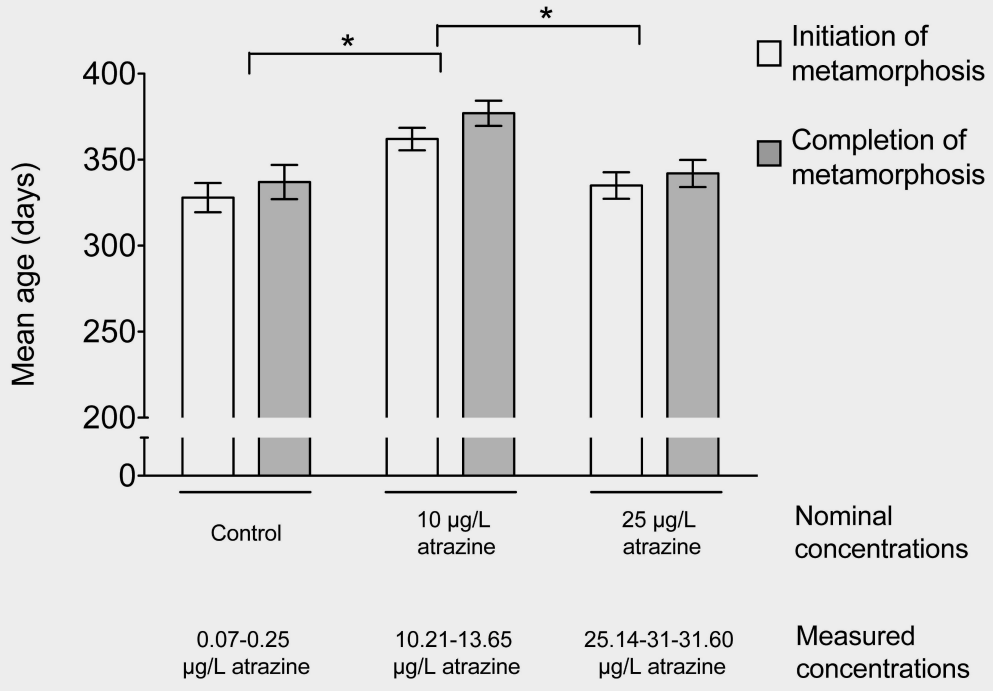

Figure 1. Effects of atrazine exposure on the timing of metamorphosis in *R. clamitans.* This is a representation of the mean age of metamorphosis initiation (white) and completion (gray) of *R. clamitans* exposed to different atrazine concentrations. Nominal and measured atrazine concentrations are listed below the x axis. Error bars represent the stardard error of the mean, * $p<0.05$. Modified from Coady *et al.*(2004).

Atrazine treatment groups did not show statistically significant differences in the other measurements (mortality, sex ratio, and gonadal development). The authors claim that the contaminated control should not affect the conclusions of the study because "there were no statistically significant, concentration-dependent effects of atrazine observed on any of the parameters investigated." (Coady *et al*. 2004, 954) This is a curious claim, since there were statistically significant differences with respect to age and size at metamorphosis between the control group and the nominal 10 µg/L concentration. See Figure 1. The authors deny that such effects were concentration-dependent because the nominal 25 µg/L treatment group did not differ from the nominal control group. I return

to the issue of controversies about the significance of such non-monotonic dose responses in toxicology in section 4.5.

Van Der Kraak *et al*. and Solomon *et al*. interpret Coady *et al*. as showing no observed effect on weight and size (Van Der Kraak et al. 2014, 19; Solomon et al. 2008, 731), no significant effect on sex ratio (Van Der Kraak et al. 2014, 21; Solomon et al. 2008, 734), and reporting no effect on the presence of testicular ovarian follicles (Van Der Kraak et al. 2014, 30). In contrast, Rohr and McCoy (2010b) claim that because the controls in Coady *et al*. were contaminated with atrazine, such claims of no effect are specious; rather than saying that there was no effect of the atrazine treatments with respect to these measurements, we should say that no such conclusions can be drawn because the putative controls contained atrazine concentrations above a plausible threshold dose for the effects under investigation (Rohr and McCoy 2010b). In short, because a compared-to-control criterion of adequacy was not met in the study, findings of no effect are invalid.

There are four replies that are open to the authors of Van Der Kraak *et al*. (2014) and Solomon *et al*. (2008). First, the authors might say that we are over-interpreting their claims of no effects. Perhaps all that is meant in saying that Coady *et al*. (2004) showed no effect of atrazine on various developmental endpoints is that there were no statistically significant differences between nominal controls and atrazine treatment groups in the study. This is a (strictly speaking, false) claim about the facts of the study rather than a claim about whether atrazine in fact has no effect on these endpoints. However, in the

weight of evidence analysis performed in Van Der Kraak *et al*., the authors code the findings of Coady *et al*. as showing "strong evidence of no effect" with respect to the question of the effect of atrazine on the time to or age at metamorphosis in amphibians. Thus, the over-interpretation reply fails.

Second, the authors might claim that they were unaware of the contaminated controls in Coady *et al*. This is implausible because the three studies share many coauthors, and Van Der Kraak *et al*. acknowledge that the controls in Coady *et al*. were contaminated (Supplemental Materials, 158). Thus, the ignorance reply fails.

Third, the authors might claim the contamination in Coady *et al*. was below the relevant threshold dose for atrazine. This reply is implausible, since there is evidence that atrazine has biological effects at the sub-ppb level (Hayes 2004; Rohr and McCoy 2010a). Additionally, Solomon *et al*. claim that the results of Saglio and Trijasse (1998) showing significant effects of atrazine on fish behavior are "impossible to interpret" in part because the controls contained atrazine at no greater than 0.235 μg/L. This suggests that Solomon et al. view the threshold dose for atrazine's biological effects at or below 0.235 μg/L.

Finally, the authors might claim that the putative compared-to-control criterion of adequacy on dose-response relationships either is not or should not be a criteria of adequacy for establishing findings of no effect. This reply is implausible, both because it

is in conflict with long-standing widely held norms in toxicology research (e.g., Hill 1965; Landis *et al*. 2003), and because, again, Solomon et al. (2008, 754) claim that the results of Saglio and Trijasse showing significant effects of atrazine on fish behavior are "impossible to interpret" in part because the controls contained atrazine at no greater than 0.235 μg/L, a claim that makes implicit appeal to a compared-to-control criterion of adequacy and a threshold dose.

Here I have used criteria of adequacy to identify problematic scientific reasoning involving the selective application of a compared-to-control criterion. When combined with principles for inferring evidence of bias, this sort of analysis in terms of criteria of adequacy can aid in separating reasonable scientific disagreement from disagreement driven by conflicts of interest.

4.4 Evidence of bias projects

In the preceding section, criteria of adequacy violations established that there were problematic interpretations of a study in a way that appeared to benefit industry. But we still need principles for inferring when conflicts of interest are likely to be the cause of the problematic scientific reasoning and practices.

Philosophers of science have only recently begun to develop principles for inferring that conflicts of interest are likely to be at work (Resnik and Elliott 2013; Elliott and Resnik

2014). There is much work to be done in establishing more principles and providing supporting reasons. Holman (2015) has developed social science methods for correlating narratives in focus groups with conflicts of interest. Bero et al. (2015) use correlations between study outcomes and funding to argue that funding is a risk factor for bias in atrazine research. Future work in social science may yield methods suitable for detecting biases via conflicts of interest in science.

In general, after having identified problematic science, claims that the science is biased by conflicts of interest or other inappropriate values are strengthened by the ability to establish 1) a motive in terms of the problematic value, 2) the mechanism by which the researchers are able to bias the research towards favored outcomes, 3) the exclusion of other possible reasons for problematic methodological or interpretive choices, and 4) a pattern of the same researcher or similarly situated researchers consistently producing favored study outcomes while other researchers in the field consistently produce contrary outcomes.

I focus here on financial conflicts of interest and propose that the strength of evidence of problematic bias via financial conflicts of interest on the part of particular researchers and particular studies increases with the increases in the number of these conditions being met. The following are a set of principles that I propose based on their intuitive plausibility and their ability to establish the four features above.

1) The scientists ignore or selectively apply well-established criteria of adequacy of their scientific community or research program (without offering defensible and publically endorsable reasons) with the outcome of producing results that are favorable to the financial interests of the scientists or their funders.

2) The scientists misrepresent the available evidence with the outcome of producing results that are favorable to the financial interests of the scientists or their funders.

3) The scientists are working in an area of research where funding is a strong predictor of study outcomes that are favorable to the financial interests of the funders.

4) The scientists fail to respond to criticisms of their work, and the failure to respond to these criticisms enables the continued production of results that are favorable to the financial interests of the scientists or their funders.

Developing new principles, justifications, and tools for providing evidence of bias is a pressing need for more effective science criticism and reforming institutions that are plagued by conflicts of interest.

4.5 The complementarity of modest demarcation projects, local criteria projects, and evidence of bias projects

I close by offering a proposal for integrating modest demarcation, local criteria, and evidence of bias approaches. I focus on the complementary roles that the three kinds of projects play in science criticism. Local criteria rather than global demarcation principles will generally be the most relevant standards for recognizing methodologically problematic science. However, modest demarcation principles, regarded as prima facie principles of good science, serve to adjudicate conflicts with regard to contested or controversial local criteria, and help to identify problematic research programs whose criteria of adequacy conflict with broadly recognized principles of good science. Evidence of bias principles will generally be most relevant to the goal of determining when methodologically problematic science is likely to be due to the illicit influence of values rather than mistakes, reasonable methodological disagreements, ignorance, *etc*.

4.5.1 Criteria of adequacy for diagnosing problematic science

As we saw in section 4.2, there is reason to think that the kinds of standards offered by global demarcation projects are liable to be subject to counterexamples and problems of vagueness. These shortcomings limit the usefulness of global demarcation principles for identifying problematic science. But as we saw at the end of Section 4.2.2 and in Section 4.3.2, local criteria of adequacy are often sufficient to identify methodologically problematic science, science that does not meet widely recognized standards among experts in the given domain.

4.5.2 Modest demarcation principles for adjudicating disputes about criteria of adequacy

Global demarcation principles, regarded as exceptionless principles applicable for all science, are generally ill-suited to pick out and explain particular instances of problematic science. However, if the most plausible principles generated by global demarcation approaches are regarded as prima facie principles, they have an important role to play in science criticism. They are helpful tools for critiquing criteria of adequacy. Here I will first give a general characterization of prima facie principles in coherence models of justification, and then illustrate their usefulness in adjudicating disagreements about criteria of adequacy.

Prima facie principles in coherence models of ethical and epistemic justification are general principles governing some domain that enjoy virtues including plausibility and argumentative support (Degrazia 1996). The principles are prima facie in the sense that we expect there to be cases in which the principle in question will have to be specified or modified in order to be preserved. Take for example, the principle that, prima facie, tests of hypothesis must be severe in order to generate good evidence in favor of that hypothesis. We saw in section 4.3.2 that there are cases in which it is plausible that combining evidence from multiple tests that are not severe can produce good evidence for a hypothesis. In other words, considerations of severity must be weighed against considerations of consilience when evaluating the quality of evidence produced by some test.[52] Considerations of consilience limit the scope of the severity principle. Here we can

[52] See also Wimsatt (2007, Chapter 4) on 'robustness.' In brief, theory or hypothesis achieves robustness when it is supported by multiple lines of evidence.

preserve a prima facie principle of severity by modifying the principle to state that in general tests of hypothesis must be severe in order to generate good evidence in favor of a hypothesis on their own, but that less severe tests can still produce good evidence in favor of a hypothesis when combined with convergent evidence yielding a consilience of inductions.

Notice that the resulting principle remains vague. We do not have an account of what exactly constitutes a severe test, what constitutes good evidence, or what constitutes a consilience of inductions. Given the death of the definitional view of concepts, we should not expect that we will be able to make the principle more precise without inviting frequent counterexample. More local domain-specific criteria of adequacy will need to do the work of setting norms for what counts as a severe test, what constitutes good evidence, or what constitutes a consilience of inductions in particular research areas.

But criteria of adequacy are frequently controversial (Brigandt 2015). More general prima facie principles of good science are often necessary to adjudicate such controversies. I will briefly give an example involving non-monotonic dose responses in toxicology.

As I noted in section 4.3.2, there is currently a debate in toxicology about how to interpret non-monotonic dose responses, especially in regulatory contexts. A non-monotonic dose response is one in which the slope of the curve describing the dose response changes signs at some point along the curve (Fagin 2012). See Figure 2.

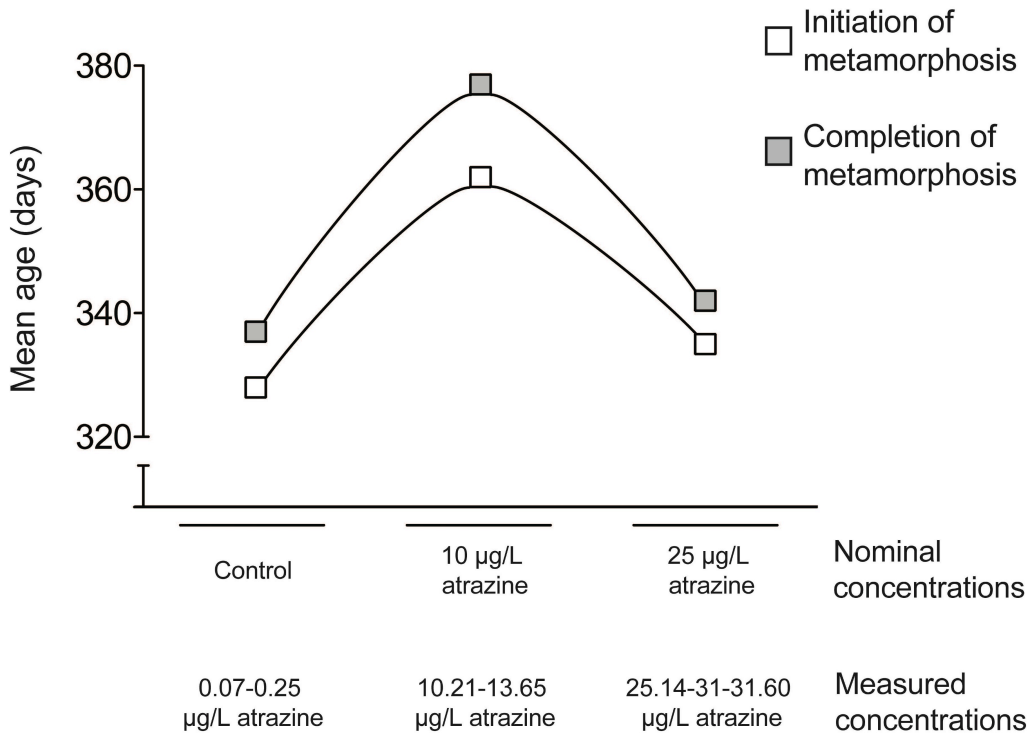

Figure 2. Non-monotonic dose response to atrazine in the timing of metamorphosis in *R. clamitans.* This is a representation of the mean age of metamorphosis initiation (white) and completion (gray) of *R. clamitans* exposed to different atrazine concentrations. These are the same results as Figure 1, but here they are represented as curves. Note that the slopes of the curves change signs at the 10 µg/L atrazine concentration. Modified from Coady *et al.*(2004).

Historically in regulatory contexts such responses have not been treated as evidence of a chemical causing an effect. In section 4.3.2, we saw Solomon *et al*. embrace this traditional perspective to claim that atrazine did not produce concentration-dependent effects on frog metamorphosis in Coady *et al*., since there were only statistically significant differences between the control group and the 10 ppb treatment group; the 25 ppb treatment group did not differ significantly from the control and thus the response was non-monotonic. In contrast, many toxicologists now accept that non-monotonic dose

responses should in general constitute evidence that a chemical is impacting a biological system (Rohr and McCoy 2010a; Fagin 2012; Elliott and Resnik 2014). So there is a conflict about a criterion of monotonicity for dose responses.

In defense of their decision to designate some non-monotonic dose responses as not relevant to questions about biological impacts of the herbicide atrazine in their review of the atrazine literature, Van Der Kraack *et al*. give up on wholesale rejection of non-monotonic dose responses in favor of counting them only when certain criteria are met. They write that non-monotonic dose responses

> can result when the response of the biological system that is the target of the chemical consists of two or more activities that might act in opposition to each other. There is no evidence to suggest that atrazine acts in this way… Therefore, non-monotonic dose responses were assigned a score of 0 [not relevant] unless a plausible mechanism… was demonstrated and there were more than two concentrations on either side of the inflection to properly characterize the response as non-monotonic. (2014, 7)

So instead of wholesale rejection, Van Der Kraack *et al*. impose two criteria on counting dose-responses as relevant evidence of an effect:

1) There must be a plausible mechanism accounting for the non-monotonicity of the curve and
2) There must be more than two concentrations on either side of the sign change in the curve describing the response.

In contrast, there is an emerging consensus among other toxicologists that non-monotonic

dose responses should be treated as evidence of an effect regardless of whether these two conditions are met (Fagin 2012; Elliott 2014). Appeals to more general principles of good science may be informative about whose perspective in this debate about criteria of adequacy is most reasonable.

We might begin with a prima facie principle, R, that scientists should consider all of the relevant evidence when reaching conclusions. This principle enjoys the virtues of plausibility and argumentative support (in the form of arguments based on considerations of traditional epistemic values like empirical adequacy and explanatory power). At first glance, Van Der Kraack *et al*. appear to be violating R by dismissing non-monotonic responses that measured fewer than seven concentrations of atrazine and/or did not provide a mechanistic interpretation of the non-monotonic response.

But Van Der Kraack *et al*. are claiming that these studies are *not* relevant, and so might deny that they are violating R. But this raises the question of why studies that demonstrate non-monotonic dose-responses that have fewer than seven measured concentrations and do not provide a mechanistic interpretation are not relevant. Van Der Kraack *et al*. have thus far not given a reason to think that dose-responses with less than seven measured concentrations are likely to be misleading with respect to non-monotonicity. They have given a reason for excluding studies without a mechanistic interpretation, but the argument is clearly invalid:

1) Some non-monotonic dose responses are caused by two or more activities that might act in opposition to each other.

2) There is no evidence to suggest that atrazine acts according to this sort of mechanism.

3) Therefore, non-monotonic dose responses for atrazine without a mechanistic interpretation are irrelevant to questions about atrazine's impact on biological systems.

This brief analysis in terms of a general principle of good science, R, does not show that the criteria for the relevance of studies generating non-monotonic dose response curves proposed by Van Der Kraack *et al*. are necessarily wrong or unreasonable. It does however place a burden of proof on Van Der Kraack *et al*. to provide more convincing reasons for why the studies the excluded are irrelevant. And the case shows how appeals to general principles of good science can be used to help adjudicate debates about criteria of adequacy. The more general principles typical of global demarcation projects (when stripped of unreasonable expectations that they will serve as exceptionless demarcation criteria covering all of science) can serve to illustrate when some positions in debates about criteria of adequacy are less reasonable than others.

4.5.3 Evidence of bias principles for inferring illicit influences of values

But local criteria of adequacy can only identify problematic scientific reasoning and methodological choices. They cannot by themselves distinguish when problematic science is likely to be due to conflicts of interest or other problematic values. Now one might wonder if this is a worthwhile goal for philosophical projects on values in science.

Steel and Powys Whyte seem to entertain the view that ascribing problematic science to the influence of e.g., conflict of interest is not important as long as problematic science is identified. They write that "an advantage of the values-in-science standard is that it makes speculations about the motives of scientists a secondary consideration. What matters first and foremost is the quality of the research…" (Steel and Powys Whyte 2012, 177)

But there are reasons to think that merely identifying research of lower quality is not sufficient given the scope of the problem of conflicts of interest and biased science. The share of research that is privately funded is large and increasing, and in domains including toxicology and pharmacology, funding source is a strong predictor of research outcome (Elliott and Resnik 2014). Reforming institutions to remediate these problems requires the development of policies, procedures, and sanctions to disincentivize science for hire. And making progress in these areas requires distinguishing science that is merely bad from science that is corrupt.

4.6 Conclusion

In conclusion, traditional global demarcation approaches for differentiating good science from problematically biased science are unlikely to meet either the theoretical aim of providing exceptionless globally applicable demarcation criteria or the practical aim of picking out science that has been corrupted by conflicts of interest or other problematic values. Instead I advocate an approach that reinterprets global demarcation principles as prima facie principles and augments these principles with more local criteria of adequacy, and principles for inferring when conflicts of interest and other problematic values are likely to be the cause of bad science.

Chapter 5: Pluralism, Monism, and Values-Lessons from Endocrine Disruption Debates

5.1 Introduction

Pluralism about key concepts is now the default position in many philosophical and scientific debates. But in many debates that are now pluralist, there were once warring monist factions. For example, species debates were once generally well-described as partisan fights over the correct definition of 'species' (e.g., Ghiselen 1987; Mayr 1987). Now many philosophers (e.g., Brigandt 2003; Ereshefsky and Reydon 2015) and scientists (e.g., Broenig et al. 2012) are more inclined to investigate the ways in which different species definitions allow for the pursuit of different epistemic and non-epistemic aims (Ludwig 2015).

In this way, the shift to pluralism has radically transformed the nature of debates where it has gained a foothold. This raises two general questions, one speculative and one normative: Which debates should we expect to become pluralistic in the future? And when is pluralism the appropriate orientation? Taylor and Vickers (2015) offer us answers; we should expect pluralism to (eventually) obtain in most or all debates about complex concepts, and pluralism is appropriate in most or all debates about complex concepts.[53]

[53] A note on terminology: in keeping with the broader philosophical and scientific literature, I am here contrasting pluralism with monism, and using 'pluralism' in roughly the same sense that Taylor and

The aim of this essay is to reject these answers and to show why the considerations that lead to them are not the only relevant ones. Taylor and Vickers' account hinges on a view about concepts, specifically the rejection of the "definitional view" of the meaning of complex concepts. But whether participants in debates will or should adopt a pluralist position depends not only on how complex concepts operate, but also on the epistemic and non-epistemic consequences of doing so. Thus, in debates in which adopting a pluralist position would have epistemic or non-epistemic consequences strongly disfavored by debate participants, we should not expect them to do so. Further, adopting a pluralist position can run counter to strong moral reasons endorsed by debate participants. In such cases, pluralism will arguably not be an appropriate position.

After introducing Taylor and Bicker's position in section 5.2, I analyze debates involving two key concept-denoting terms in endocrine disruption research. In section 5.3, I will examine debates about 'potency.' Multiple definitions of 'potency' play distinct and necessary roles endocrine disruption research, and appealing to multiple definitions of 'potency' is consonant with the epistemic and non-epistemic aims of the parties in the debate. So debate participants are 'potency' pluralists. But in many of the same journal articles in which these debate participants are pluralists about 'potency,' they remain steadfast monists about 'endocrine disruptor.' In 5.4, I argue that this is a predictable and arguably appropriate response because of the epistemic and non-epistemic costs

---

Vickers use 'fragmentation.' They contrast pluralism with eliminativism, and fragmentation with monism. Since I will not discuss eliminativism, monism is the appropriate contrast to pluralism in this essay.

associated with being pluralist about 'endocrine disruptor.'

I close by suggesting that my conclusion that pluralism is not always to be expected or recommended is consonant with the "pluralist stance" articulated by Kellert et al. (2006). The pluralist stance is not committed to pluralism obtaining in all areas of science; "the pluralism advocated is local rather than universal." (Kellert et al. 2006, xxiii) Instead, the stance requires investigation of particular areas of science with respect to whether multiple representational and explanatory schemes are present, and what is motivating the adoption of those schemes.

5.2 Taylor and Vickers' Prediction and Promotion of Pluralism

5.2.1 Prediction

Say that *pluralism* about a term T obtains in a debate when

a) T has more than one candidate definition and
b) More than one of T's candidate definitions are employed for distinct roles in discourse and
c) Debate participants generally acknowledge that no one candidate definition of T is the only correct definition.

Say that *monism* about a concept-denoting term T obtains when (a) and (b) are true, but not (c).[54]

[54] I will not here consider debates in which (a) or (b) are false, since (a) and (b) being true are necessary for the kinds of conceptual debates that typically occupy philosophers and scientists. The criteria (a)-(c) are adapted from Taylor and Vicker's criteria for 'fragmentation.'

Taylor and Vickers offer an account of why we should expect pluralism to obtain in most or all debates about complex concepts. The account begins with the rejection of a traditional view about the meaning of concepts. Traditionally terms denoting complex concepts have been thought of as definable by a set of properties that are individually necessary and jointly sufficient to capture the meaning of a concept-denoting term and its extension. Call this the definitional view of the meaning of concept-terms.

In monistic conceptual debates, definitions die by counterexample. Suppose the definition of 'species' is a population of organisms that can successfully interbreed. This definition aims to capture the meaning of 'species' and to set conditions for what properly counts as a species. Counterexamples are cases that either 1) meet the definition, but do not intuitively count as instantiations of the concept, or 2) do not meet the definition, but intuitively count as instantiations of the concept. Call this the *extensional adequacy criterion* (EAC) on definitions (Taylor and Vickers 2015). Grey wolves and domestic dogs can successfully interbreed, but they have morphological and ecological differences that suggest intuitively that they are different species. Likewise, New Mexico whiptail lizards comprise a tax on that intuitively seem like a species, but they only reproduce asexually, and so do not interbreed. These counterexamples show that the interbreeding definition of 'species' does not meet the EAC.

According to Taylor and Vickers, after a period of inevitable failure (see section 4.2) interlocutors in conceptual debates come to recognize that no definitions of the concept

under debate can survive the EAC, but that many of the definitions are useful for different purposes. For example, Ludwig (2015) highlights the impressive predictive and explanatory achievements of the "interbreeding" species definition, while also noting that the definition is subject to counterexample.

On Taylor and Vickers' account, conceptual debate participants come to recognize the usefulness of definitions rather than the EAC as the relevant standard for judging definitions. Debate participants soon find that more than one definition is useful, since concept-denoting terms are marshalled for a variety of uses.

5.2.2 Promotion

Even if pluralism is eventually inevitable, Taylor and Vickers urge participants in conceptual debates to adopt pluralism sooner rather than later. Given the "death" of the definitional view of concepts, they argue that any recalcitrant monists owe an account of why the debates they are monists about do not fit with the predictive account offered in subsection 2.1. That is, the monist should have to say why we should expect their pet concept to survive the EAC, or what the motivation for having a single correct definition is given that it cannot be extensional adequacy.

In section 5.4, I aim to develop another possible response for at least some recalcitrant monists; there are sometimes good epistemic and non-epistemic reasons to resist

pluralism. But first, in section 5.3, I will introduce a conceptual debate that seems to fit well with Taylor and Vickers' account. The respective analyses in sections 5.3 and 5.4 will highlight the contrast between cases where epistemic and non-epistemic considerations (considered to be important by the parties in the debate) do (5.3) and do not (5.4) favor pluralism.

5.3 Potency Pluralism in Endocrine Disruption Debates

5.3.1 Parties and interests in recent endocrine disruption debates

Over the past several years, different groups of researchers have offered strikingly different conclusions about the risks posed by endocrine disrupting chemicals. In 2012, the World Health Organization (WHO) and United Nations Environment Program (UNEP) published a "state of the science" review of endocrine disrupting chemicals (EDCs). The WHO/UNEP report found that there was sufficient evidence to conclude that EDCs pose significant health risks to humans and wildlife (WHO/UNEP 2012). The publication of the report initiated a series of heated exchanges between the study's authors and supporters and a skeptical group of researchers who argue that WHO/UNEP report does not present a balanced picture of the uncertainties associated with EDCs and thus does not represent a true "state of the science" review (Autrup et al. 2015; Bergman et al. 2015). In addition to disagreements about sufficiency of evidence for hypotheses involving EDCs and criteria governing systematic reviews of EDC literature, exchanges

between supporters and skeptics of the WHO/UNEP report highlight persistent disagreements involving key concepts in EDC research including 'endocrine disruptor,' 'endocrine function,' 'adverse effect,' 'potency,' and 'threshold.'

Both support for and skepticism about the WHO/UNEP report tends to be correlated with disciplinary background and funding. The group of authors that I will label "Skeptics" tend to have backgrounds in traditional toxicology[55] and pharmacology and tend to receive a large portion of their funding from the chemical industry. Representatives of the Skeptics include Borgert et al. (2013), Dietrich et al. (2013), Nohynek et al. (2013) Lamb et al. (2014; 2015), and Autrup et al. (2015). The group of authors that I will label "Supporters" tend to have backgrounds in endocrinology and tend to receive their funding from governments and NGOs rather than private industry. Representatives of the Supporters include Zoeller et al. (2014) and Bergman et al. (2013; 2015), and Bourguignon et al. (2015).

Endocrine disruptor research was founded on the idea that traditional toxicological methods were failing to detect some of the harmful effects of chemicals (Colborn et al. 1993; Krimsky 2000). Consonant with this founding idea of their research program, the main general normative epistemic claim of Supporters is that traditional toxicological

[55] Toxicology is a heterogeneous discipline with a large variety of classificatory and explanatory aims and methods. The perspective in toxicology that I am referring to with the term, 'traditional toxicology,' has the following features: an investigative emphasis on acute toxicity and interruption of homeostasis, a methodological assumption of linear dose responses with thresholds, and methodological commitment to granting epistemic and regulatory privilege to experiments that use a relatively small set of "validated" test procedures and biological endpoints and follow standardized reporting requirements. My use of 'traditional toxicology' corresponds roughly to Bourguignon et al. (2015)'s 'non-endocrine perspective.'

tests and methodological standards, and current regulatory tests and standards for endocrine disruptors, tend to focus on an inappropriately limited set of biological endpoints under an inappropriately limited set of conditions. They claim that toxicological research on endocrine disruptors should be reformed in light of "the principles of endocrinology." (Bourguignon et al. 2015) According to these principles, the effects of modulations of hormone systems will be highly context-dependent, and thus hard to detect with tests that only investigate only a very small subset of these contexts. Thus, standards of evidence for EDC effects, and the kinds of tests employed to detect these effects should reflect the highly context-specific nature of the effects. Supporters tend to be explicit about their related non-epistemic goal of preventing harms to public health and the environment by limiting harmful chemical exposures. For example, Bourguignon et al. (2015) write that, "The public health consequences of either inaction or regulatory decisions missing the endocrine principles would be worrying, possibly for several generations." (10)

In contrast, Skeptics tend to claim that "chemicals with hormonal activity can be subjected to the well-evaluated health risk characterization approach used for many years" in traditional toxicology (Autrup et al. 2014). Skeptics are also often forthright about their non-epistemic goal of preventing burdensome or unnecessary regulation based upon what they see as the overly permissive epistemic standards, and overly strict regulatory ambitions of Supporters. For example, Dietrich et al. (2013) write that "Regulations that profoundly affect human activities... should not be based on irrelevant tests…". (2) Additionally, Skeptics often express worries that the approach favored by

Supporters would lead to an unmanageably large number of chemicals being treated as potentially harmful, and thus make it difficult to set regulatory priorities. Nohynek et al. (2013) write that,

> Overall, the entire discussion whether man-made chemicals with hormone-like activity may pose a risk to human health has a paradoxical aspect: if such activities, however small, could actually pose a potential health risk, then it would make sense to worry about all substances that possess such activities, particularly when potent oestrogens… are present in human food, such as phytoestrogens. (301)

In this subsection I have introduced Supporters and Skeptics of the WHO/UNEP (2012) report, and summarized some of their central epistemic and non-epistemic motivations. In the following sub-sections, I will argue that adopting a pluralist attitude towards 'potency' is consistent with these motivations.

5.3.2 Two characterizations of 'potency'

Although conceptual debates involving 'potency' are central in the endocrine disruptor controversy, consonant with Taylor and Vickers' account, both Supporters and Skeptics tend to be pluralists about the concept of 'potency.' While there are many different meanings of 'potency' at work in EDC debates, for the sake of simplicity I will focus only on two. Although the two are not formally defined in the manner of a philosophical definition, they are clearly characterized and differentiated in the literature as follows:

Cellular/receptor level potency: "Together, affinity and efficacy determine the potency of a ligand [an agent] to activate specific hormone receptors and to elicit specific cellular responses in target tissues." (Borgert et al., 2013; Lamb et al. 2014)[56]

Organism level potency: "At the whole organism level, potency relates to the ability of a substance to produce a biological effect and may be substantially different from the potency measured with *in vitro* assays." (Lamb et al. 2014)

Both Supporters and Skeptics explicitly use both definitions of 'potency,' and there is a plausible history of the term 'potency' that fits Taylor and Vickers story about how terms come to have different meanings. Taylor and Vickers write that,

> expressions get introduced into a field… at a time when only a small fraction of the interesting phenomena have been identified. As any field develops the phenomena to be explained multiply. The new phenomena demand explanation, and the terms which have already been involved in successful explanations are invoked again for the new, related phenomena. (22)

'Potency' has been a term in toxicology and pharmacology since before there were experiments on and experimentally-informed theories about the cellular/receptor level (Borgert et al. 2013). When the capacity of chemicals to produce cellular and receptor level effects became an experimentally-accessible phenomenon, it appears the old term was employed with a new meaning.

5.3.3 Pluralism about 'potency' is consistent with the epistemic and non-epistemic

[56] 'Affinity' refers to the propensity of a chemical (ligand) to bind to a cellular hormone receptor. "Efficacy" refers to the ability of the resulting ligand-receptor complex to elicit a cellular response. The cellular responses of interest are often changes in gene expression.

interests of Supporters and Skeptics

We saw in section 5.3.1 that Supporters and Skeptics have opposing epistemic and non-epistemic aims. However, pluralism about potency is consistent with these aims. This is true because, first, the research methodologies and investigative strategies recommended by both Supporters and Skeptics involve characterizing the potency of chemicals at both the cellular/receptor and organismal levels (Lamb et al. 2014; Zoeller et al. 2014).

Second, the availability of multiple meanings of 'potency' allow for both Supporters and Skeptics to develop rhetoric in favor of their views and against those of their debate opponents. For example, Skeptics point to the relatively low cellular/receptor potency of benzylparaben in yeast reporter-gene assays to imply that benzylparaben, and other parabens, are unlikely to have high organism-level potency (Nohynek et al. 2013). Skeptics use this argument to dismiss the claims of Supporters that parabens are potentially dangerous endocrine disruptors. Likewise, Supporters use the ambiguity of 'potency' to argue that "More potency can have less effect." (Bourguignon et al. 2015) This argument is used to portray the Skeptics' view of endocrine systems as overly simplistic.

These strategies, where 'potency' is used in subtly different ways throughout the course of arguments, create rhetoric conducive to the respective aims of both Supporter and Skeptics. Pluralism about 'potency' is conducive to promoting the research agendas and non-epistemic aims of both Supporters and Skeptics. In the next section, I present a case where pluralism runs counter to these agendas and aims.

5.4 'Endocrine Disruptor' Monism

5.4.1 Two definitions of 'endocrine disruptor'

There have been many definitions of 'endocrine disruptor' proposed over the past 20 years (Zoeller et al., 2014). However, the two major candidate definitions that divide Supporters and Skeptics are as follows:

WHO/ICPS (WI): An exogenous substance or mixture that alters function(s) of the endocrine system and consequently causes adverse effects in an intact organism, or its progeny, or (sub)populations. [57]

Endocrine Society (ES): An exogenous chemical, or mixture of chemicals, that interferes with any aspect of hormone action.

WI is much more demanding in the sense that it requires that a chemical or mixture be demonstrated to have more properties in order to be labeled as an endocrine disruptor than does the **ES**. **WI** requires that it be demonstrated that

1) The chemical or mixture alters endocrine function
2) The chemical or mixture causes an adverse health effect (at the organism or population level) and
3) The adverse health effect is caused via an endocrine altering mode of action

[57] This definition is drawn from a 2002 WHO report (Damstra et al.), not the 2012 WHO report (Bergman et al.) about which Skeptics are skeptical.

ES, in contrast, is much more permissive with respect to designating a chemical as an EDC. It merely requires that a chemical be demonstrated to have some effect on some aspect of hormone action.

5.4.2 Goals, values, and definitional commitments

Skeptics embrace the **WI** but deny the **ES**. The opposite is true for Supporters.[58] It is not difficult to see why this is the case. Recall that Skeptics are committed to i) a research program that centers on traditional toxicological testing procedures and methodological commitments, ii) the avoidance of unnecessary regulation, and iii) avoiding situations in which it is difficult to set regulatory priorities because there are an unmanageably large number of chemicals that are potential objects of regulation.

With respect to i), Skeptics argue that **WI** requires "dose–incidence curve for adverse effects in intact animals from appropriate toxicity studies" to establish a no-observed-adverse-effect level or benchmark dose. (Autrup et al. 2014, 3) These requirements suggest traditional toxicological tests and standards. With respect to ii) and iii), Lamb et al. (2014) argues that the WHO/UNEP report inappropriately draws conclusions that various chemicals are endocrine disruptors without having met the three **WI** criteria (29), and expresses concerns about the likely regulatory effects of this aspect of the report (35).

Recall that Supporters are committed to iv) reforming endocrine disruption research

[58] Although some Supporters (e.g., Zoeller et al. 2014) claim to accept the WHO/ICPS definition, they insist on interpreting the WHO/ICPS definition such that it is identical to the Endocrine Society definition. I will therefore treat these Supporters as embracing the Endocrine Society definition and rejecting the WHO/ICPS definition.

standards to reflect the complexity of the endocrine system and its effects, and v) preventing environmental and public health harms by limiting EDC exposures. With respect to iv) Supporters argue that **ES** is the appropriate definition of 'endocrine disruptor' for a suitably reformed EDC research agenda (Zoeller et al. 2012; 2014). With respect to v) Bourguignon et al. (2015) support the adoption of **ES** as an element of the reforms that they see as necessary to safeguard human health from chemicals that may be impacting the timing of the onset of puberty in human populations.

5.4.3 Reply to an objection

Recall from section 5.2.1 that the criterion separating pluralists from monists is (c): Debate participants generally acknowledge that no one candidate definition of T is the only correct definition. Recall from section 5.4.1 that I acknowledged that there have been many 'endocrine disruptor' definitions over the years. Some of these definitions have been made part of regulations, for example in the USEPA Endocrine Disruptor Screening Program. Both Supporters and Skeptics have used various definitions of 'endocrine disruptor' in their published work in order to make logical contact with regulatory definitions.

Thus, one might argue that (c) obtains in debates between Supporters and Skeptics involving 'endocrine disruptor.' Supporters embrace **ES** and deny **WI**, and Skeptics embrace **WI** and deny **ES**, but this does not, so the objection goes, make them monists. It merely means that they are pluralists who respectively deny one definition of 'endocrine

disruptor,' but are not committed to the claim that their favored definition is the only correct one.

In reply, I point to a strategy used by both Supporters and Skeptics for interpreting regulatory definitions. This is the strategy of interpreting regulatory decisions in ways that are consistent with their favored definition. For example, Zoeller et al. (2014) write that, presumably regrettably, "The various definitions of an EDC proposed by regulatory agencies are not likely to change." But they nonetheless argue for interpreting various definitions as equivalent to **ES**. While both Supporters and Skeptics have used several regulatory definitions in their discussions for expediency, they systematically refuse legitimacy to interpretations of those definitions that are inconsistent with their own favored definition. In this sense (c) fails, and Supporters and Skeptics are properly described as monists.

5.4.4 A defense of monism against Taylor and Vickers predictive and normative pluralism

Taylor and Vickers account predicts that pluralism will (eventually) come to be the position adopted by parties in all or most debates that centrally feature complex concepts. Their predictive view is based on the general inability of candidate definitions of concept-denoting terms to survive the EAC. After parties in these debates realize that all candidate definitions fail by the EAC, they will come to judge definitions by their usefulness rather than the EAC. When parties recognize that more than one definition is useful, because the situations in which the term is deployed are diverse, pluralism

follows.

As we saw in section 5.3, this account fits well with the pluralist attitudes towards 'potency' that we see in endocrine disruptor debates. 'Potency' has multiple distinct definitions, and different meanings are appropriate to different purposes. Additionally, there is a plausible story to tell, consonant with Taylor and Vickers's account, about how 'potency' came to have multiple distinct and useful meanings (5.3.2). But importantly, as shown in 5.3.3, pluralism about 'potency' is also consistent with the goals and interests of the debating parties.

But, contrary to Taylor and Vickers prediction of pluralism eventually obtaining in most or all debates centrally involving complex concepts, we should not expect pluralism to obtain when pluralism runs strongly counter to the goals and interests of the parties in the debate. In sections 5.4.2 and 5.4.3, I argued that Supporters and Skeptics are monists about 'endocrine disruptor,' and described how their respective commitments to privileged definitions of 'endocrine disruptor' fit with their preferred research orientations and their nonepistemic goals.

But why would being a pluralist about 'endocrine disruptor' not be consonant with the respective epistemic and non-epistemic goals of Supporters and Skeptics? Suppose that a Skeptic adopted a pluralist position about 'endocrine disruptor.' This would presumably involve the Skeptic assenting to something like the following: **WI** is a useful definition of 'endocrine disruptor' for studying EDCs within a traditional toxicological framework and avoiding burdensome and potentially unnecessary regulation. However, if your epistemic

goals and standards are like those of the Supporters, and you are willing to risk potentially unnecessary and burdensome regulations to protect against potential human health and environmental threats about which there is a great deal of uncertainty, then **ES** is an appropriate definition. But neither **WI** nor **ES** is to be privileged as the correct definition of 'endocrine disruptor,' because both are useful for different purposes.

There is a sense then in which, by being a pluralist about 'endocrine disruptor,' the pluralist Skeptic would be acknowledging the equal legitimacy of the epistemic goals and standards, and the non-epistemic aims of Supporters. But these are precisely what the Skeptic wants to deny. An analogous account can be described for a pluralist Supporter. We can generalize and say that insofar as being a pluralist about a concept-denoting term would involve granting equal legitimacy to epistemic and non-epistemic goals that debating parties are at great pains to deny the equal legitimacy of, we should not expect those parties to be pluralists. It in failing to acknowledge this feature of choices between monism and pluralism that Taylor and Vickers' predictions about the spread of pluralism miss the mark.

But what about their normative promotion of pluralism? Taylor and Vickers offer a challenge to recalcitrant monists to defend their monism in the face of the death of the definitional view of concepts. The monist must either say that the concept about which they are a monist can survive the EAC, or they must give a reason other than extensional adequacy for preferring a single definition (section 5.2.2). We are now in a position to meet this challenge. The recalcitrant monist is entitled to remain a monist to the extent that being a pluralist would require the legitimizing of epistemic or non-epistemic goals,

standards, or values whose legitimacy the monist denies.

To illustrate in the 'endocrine disruptor' case, the term 'endocrine disruptor' is widely acknowledged to be a powerful term (Krimsky 2002). The term implies a threat to the normal functioning of biological bodies (Borgert et al. 2013). Additionally, the term has served to unify in the minds of the public a set of previously unrelated worrisome phenomena, including diverse negative impacts to humans and wildlife (Krimsky 2002; Elliott 2011). Pluralism on the part of either Supporters or Skeptics would be ceding some of the power to control the use of this term to their dialectical opponents. Both Supporters and Skeptics can offer moral reasons against doing so. In the case of Supporters, granting equal legitimacy to the **WI** definition would give Skeptics more leeway to define a broad class of potentially dangerous chemicals as non-endocrine disruptors, and thereby increase the likelihood that these chemicals will be under-regulated. From the perspective of a Supporter, this would presumably be morally wrong. And while I am less convinced of its cogency, Skeptics also offer a moral argument for a more restricted view of what counts as an endocrine disruptor (Deitrich et al. 2013). If being a pluralist requires acting in ways contrary to one's considered moral judgments, so much the worse for pluralism.

I would like to note in passing that Taylor and Vickers' descriptions of several debates as pluralistic begin to look suspect in light of the analysis that I have given here. While some debates that prominently feature concept-denoting terms like 'race' and 'intelligence' might well be pluralistic in character (Taylor and Vickers 2015), other debates involving these terms appear to be debates between recalcitrant monists (Ludwig

2015). Maybe with good reason.

5.4.5 Monism and the pluralist stance

Perhaps surprisingly, the defense of monism that I have presented in section 5.4.4 is arguably more consonant with an influential pluralist position than is the pluralism of Taylor and Vickers. Kellert et al. (2006) articulate a "pluralist stance" which rejects *a priori* commitments to the view that science's goal is a single unified representation of the natural world. Like Taylor and Vickers account, the pluralist stance rejects the EAC as the primary criterion of conceptual analysis; rather than the ability to survive counterexample, the pluralist stance holds that "[c]onceptual analyses ought to be evaluated on the basis of what they help us understand and investigate." (Kellert et al. 2006, xxvi) But in contrast to Taylor and Vickers' account, which recommends pluralism *a priori* on the basis of considerations about the nature of concepts, the pluralist stance is not committed to recommending pluralism in all debates about complex scientific concepts. Instead, the pluralist stance recommends the study of particular areas of and episodes in science with respect to whether multiple representational and explanatory schemes are present, and with respect to possible motivations for and successes of the adoption of multiple representational and explanatory schemes. In this sense, and again in contrast to Taylor and Vickers' account, the pluralist stance recommends context-specific local pluralisms based on investigations of particular cases, rather than universal pluralism based on *a priori* commitments.

Advocates of the pluralist stance are committed in general to the possibility that some

areas of scientific research may be properly monist in character. That is, it is possible that some areas of science may be well-characterized as providing a single successful and consistent account of their domain of inquiry utilizing a single representational and explanatory scheme. But this general commitment does not touch directly on the issues at stake here; 'endocrine disruptor' remains contested, and it does not appear likely that the representational and explanatory schemes of either Supporters or Skeptics will hold successful monistic sway over endocrine disruption research in the near future. So, if the general commitment to the possibility of successful monism is not the reason, how and why might the pluralist stance be consonant with defense of monism that I have articulated here?

The account of 'pluralism' and 'monism' articulated by the pluralist stance is different than the account of 'pluralism' and 'monism' given in section 5.2.1 and used thus far here. Instead the pluralism of the pluralist stance is defined by the rejection of *a priori* commitments to meta-scientific monism. This meta-scientific monism is roughly the view that the goal of science is a single unified account of the natural world expressed in a single consistent representational and explanatory scheme. The rejection of meta-scientific monism is thus consistent with particular scientists in particular domains being monists about particular concepts. The pluralist stance merely entails that meta-scientific analyses not commit themselves *a priori* to a monistic criterion of successful science whereby a science is successful to the extent that it provides a univocal account of its domain of inquiry.

In summary, the pluralist stance does not require that scientists in all domains themselves

become pluralists about key concepts. And in rejecting *a priori* commitments to monism and pluralism, but instead focusing on the investigation of particular cases, the defense of monism that I have given here seems closer to the program of science analysis advocated by Kellert et al. than is the *a priori* promotion of pluralism advocated by Taylor and Vickers.

Bibliography

Anderson, Elizabeth (2004). Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce. *Hypatia* 19(1): 1–24.

Anderton, D. L., Anderson, A. B., Oakes, J. M., & Fraser, M. R. (1994). Environmental equity: the demographics of dumping. *Demography*, *31*(2), 229-248.

Appleby, A. P. (2005). A history of weed control in the United States and Canada-a sequel. *Weed Science*, *53*(6), 762-768.

Autrup, H., Barile, F. A., Blaauboer, B. J., Degen, G. H., Dekant, W., Dietrich, D., ... & Kacew, S. (2015). Principles of pharmacology and toxicology also govern effects of chemicals on the endocrine system. *Toxicological Sciences*, kfv082.

Aviv, Rachel. (2014). A Valuable Reputation. *The New Yorker*, February 10.

Barringer, Felicity (2008). Hermaphrodite Frogs Found in Suburban Ponds. *The New York Times*, April 4.

Bausman, W. (2015). *Neutral theory, biased world* (Unpublished doctoral dissertation). University of Minnesota.

Bechtel, W. (1993). Integrating sciences by creating new disciplines: The case of cell biology. *Biology and Philosophy*, *8*(3), 277-299.

Bergman, Å., Heindel, J., Jobling, S., Kidd, K., & Zoeller, R. T. (2012). State-of-the-science of endocrine disrupting chemicals, 2012. *Toxicology Letters*, *211*, S3.

Bero, L., Anglemyer, A., Vesterinen, H., & Krauth, D. (2015). The relationship between study sponsorship, risks of bias, and research outcomes in atrazine exposure studies conducted in non-human animals: Systematic review and meta-analysis. *Environment international*.

Biddle, Justin B. (2013). State of the Field: Transient Underdetermination and Values in Science. *Studies in History and Philosophy of Science* 44: 124–33.

Biddle, Justin B. (2016). Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease. *Perspectives on Science* 24, 192–205.

Birke, Lynda (2000). Sitting on the Fence: Biology, Feminism and Gender-Bending Environments. *Women's Studies International Forum* 23(5): 587–99.

Boenigk, J., Ereshefsky, M., Hoef-Emden, K., Mallet, J., & Bass, D. (2012). Concepts in protistology: species definitions and boundaries. *European Journal of Protistology*, *48*(2), 96-102.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, 303-352.

Boone, M. D., Bishop, C. A., Boswell, L. A., Brodman, R. D., Burger, J., Davidson, C., ... & Weir, S. (2014). Pesticide regulation amid the influence of industry. *BioScience*, biu138.

Boone, M. D., Bishop, C. A., Boswell, L. A., Brodman, R. D., Burger, J., Davidson, C., ... & Rohr, J. R. (2014). Pesticide regulation amid the influence of industry. *BioScience*, biu138.

Borgert, C. J., Baker, S. P., & Matthews, J. C. (2013). Potency matters: thresholds govern endocrine activity. *Regulatory Toxicology and Pharmacology*, *67*(1), 83-88.

Bourguignon, J. P., Juul, A., Franssen, D., Fudvoye, J., Pinson, A., & Parent, A. S. (2016). Contribution of the Endocrine Perspective in the Evaluation of Endocrine Disrupting Chemical Effects: The Case Study of Pubertal Timing. *Hormone research in paediatrics*.

Brigandt, I. (2003). Species pluralism does not imply species eliminativism. *Philosophy of Science*, *70*(5), 1305-1316.

Brigandt, Ingo (2010), "Beyond reduction and pluralism: Toward an epistemology of explanatory integration in biology", *Erkenn* 73:295–311

Brigandt, I. (2015). Social values influence the adequacy conditions of scientific theories: beyond inductive risk. *Canadian Journal of Philosophy*, *45*(3), 326-356.

Brigandt, Ingo and Love, Alan (2012). Reductionism in Biology, *The Stanford Encyclopedia of Philosophy (Summer 2012 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2012/entries/reduction-biology/>.

Brown, Matthew (2013) Values in Science Beyond Underdetermination and Inductive Risk. *Philosophy of Science* 80(5), 829–39.

Burian, R. M. (1997). Exploratory experimentation and the role of histochemical techniques in the work of Jean Brachet, 1938-1952. *History and Philosophy of the Life Sciences*, 27-45.

Caraty, A., Locatelli, A., & Martin, G. B. (1989). Biphasic response in the secretion of gonadotrophin-releasing hormone in ovariectomized ewes injected with oestradiol. *Journal of endocrinology*, *123*(3), 375-382.

Carr, J. A., Gentles, A., Smith, E. E., Goleman, W. L., Urquidi, L. J., Thuett, K., ... & Van Der Kraak, G. (2003). Response of larval Xenopus laevis to atrazine: assessment of growth, metamorphosis, and gonadal and laryngeal morphology. *Environmental Toxicology and Chemistry*, *22*(2), 396-405.

Chester-Jones, I., Ingleton, P. M., & Phillips, J. G. (Eds.). (2013). *Fundamentals of comparative vertebrate endocrinology*. Springer Science & Business Media.

Churchman, C. West. (1948) Statistics, Pragmatics, Induction. *Philosophy of Science* 15(3): 249–68.

Coady, K. K., Murphy, M. B., Villeneuve, D. L., Hecker, M., Jones, P. D., Carr, J. A., ... & Giesy, J. P. (2005). Effects of atrazine on metamorphosis, growth, laryngeal and gonadal development, aromatase activity, and sex steroid concentrations in Xenopus laevis. *Ecotoxicology and environmental safety*, *62*(2), 160-173.

Coady, K., Murphy, M., Villeneuve, D., Hecker, M., Jones, P., Carr, J., ... & Giesy, J. (2004). Effects of atrazine on metamorphosis, growth, and gonadal development in the green frog (Rana clamitans). *Journal of Toxicology and Environmental Health*, *67*(12), 941-957.

Colborn, T., vom Saal, F. S., & Soto, A. M. (1993). Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environmental health perspectives*, *101*(5), 378.

Colborn, Theo, Dianne Dumanoski, and John Peter Meyers. (1996). *Our Stolen Future: Are We Threatening Our Fertility, Intelligence and Survival? A Scientific Detective Story*. New York: Dutton. Columbia University Press.

Craver, C. F. (2007). *Explaining the brain*. Oxford University Press.

Damstra, T., Barlow, S., Bergman, A., Kavlock, R., & Van Der Kraak, G. (2002). Global assessment of the state-of-the-science of endocrine disruptors. *Geneva: World Health Organization*.

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science*, 43-64.

DeGrazia, D. (1996). *Taking animals seriously: mental life and moral status*. Cambridge University Press.

Di Chiro, Giovanna (2010). Polluted Politics? Confronting Toxic Discourse, Sex Panic, and Eco-Normativity." In *Queer Ecologies: Sex, Nature, Politics, Desire,* edited by Catriona Mortimer-Sandilands and Bruce Erickson, 199–230. Indianapolis: Indiana University Press.

Dietrich, D. R., Aulock, S. V., Marquardt, H., Blaauboer, B., Dekant, W., Kehrer, J., ... & Lang, F. (2013). Scientifically unfounded precaution drives European Commission's recommendations on EDC regulation, while defying common sense, well-established science and risk assessment principles. *Chem Biol Interact*, *205*(1), A1-5.

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

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science* 67(4): 559–79.

Douglas, H. (1998) The Use of Science in Policy-Making: A Study of Values in Dioxin Science." PhD diss, University of Pittsburgh.

Dupré, J. (1993). *The disorder of things: metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.

Elliott, Kevin C. (2009). Ethical Significance of Language in the Environmental Sciences: Case Studies from Pollution Research. *Ethics, Place & Environment: A Journal of Philosophy & Geography* 12(2): 157–73.

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

Elliott, K. C., & McKaughan, D. J. (2009). How values in scientific discovery and pursuit alter theory appraisal. *Philosophy of Science*, *76*(5), 598-611.

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

Elliott, K. C., & Resnik, D. B. (2014). Science, policy, and the transparency of values. *Environmental Health Perspectives (Online)*, *122*(7), 647.

Elliott, Kevin C., and David Willmes (2013). Cognitive Attitudes and Values in Science. *Philosophy of Science* 80(5): 807–17.

Ereshefsky, M., & Reydon, T. A. (2015). Scientific kinds. *Philosophical Studies*, *172*(4), 969-986.

Fagin, D. (2012). The learning curve. *Nature*, *490*(7421), 462-465.

Fausto-Sterling, Anne (2000). The Five Sexes, Revisited." *The Sciences* 40(4): 18–23.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1-35.

Frigg, R. (2006). Scientific representation and the semantic view of theories. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *21*(1), 49-65.

Ghiselin, M. T. (1987). Species concepts, individuality, and objectivity. *Biology and Philosophy*, *2*(2), 127-143.

Gutierres JB, Teem Jl (2006). A model describing the effect of sex-reversed YY ␣sh in an established wild population: The use of a Trojan Y chromosome to cause extinction of an introduced exotic species. *J Theor Biol* 241:333–341.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science* (Vol. 5, No. 1). Cambridge: Cambridge University Press.

Halina, Marta (2015). There Is No Special Problem of Mind-Reading in Non-Human Animals. *Philosophy of Science* 82(3): 473–90.

Hansson, Sven Ove (2015). Science and Pseudo-Science. *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2015/entries/pseudo-science/>.

Hayes *et al*. (2010). Atrazine induces complete feminization and chemical castration in male African clawed frogs(Xenopus laevis) *PNAS* March 9, 2010 vol. 107 no. 10 4612-4617

Hayes *et a*l. (2011). Demasculinization and feminization of male gonads by atrazine: consistent effects across vertebrate classes. *J Steroid Biochem Mol Biol* 127:64 –73

Hayes, T. B. (2004). There is no denying this: defusing the confusion about atrazine. *Bioscience*, *54*(12), 1138-1149.

Hayes, T. B. (2004). There is no denying this: defusing the confusion about atrazine. *Bioscience*, *54*(12), 1138-1149.

Hayes, Tyrone (2005). Welcome to the revolution: Integrative Biology and Assessing the Impact of Endocrine Disruptors on Environmental and Public Health. *Integr. Comp. Biol*., 45:321–329

Hempel, Carl G. (1954). A Logical Appraisal of Operationalism. In *The Validation of Scientific Theories*, edited by Philipp G. Frank, 52–67. Boston: The Beacon Press.

Hempel, Carl G. (1958). The Theoretician's Dilemma: A Study in the Logic of Theory Construction. In *Concepts, Theories, and the Mind-Body Problem*, edited by Herbert Feigl, Michael Scriven and Grover Maxwell, 37–97. Minneapolis: University of Minnesota Press.

Hempel, Carl G. (1965). Science and Human Values. In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press: 81–96.

Hempel, Carl G. (1981). Turns in the Evolution of the Problem of Induction. *Synthese* 46(3): 389–404.

Herbers, John M. (2007). Watch Your Language! Racially Loaded Metaphors in Scientific Research. *BioScience* 57(2): 104–5.

Hill, A. B. (1965). The environment and disease: association or causation?. *Proceedings of the Royal Society of Medicine*, *58*(5), 295.

Holman, Bennett (2015). Asymmetric Epistemic Arms Races and the Prospects of Collaborative Research. Paper presented at the Collaboration Conundrum Conference, University of Notre Dame.

Huelgas-Morales, Gabriela and Jack Powers (2016) Adding Conceptual Analysis to the Experimentalist's Toolkit. Manuscript in preparation.

Jablonowski, N. D., Schäffer, A., & Burauel, P. (2011). Still present after all these years: persistence plus potential toxicity raise questions about the use of atrazine. *Environmental Science and Pollution Research*, *18*(2), 328-331.

James, William (1896) The Will to Believe. *The New World* 5: 327–47.

Jeffrey, Richard C. (1956). Valuation and Acceptance of Scientific Hypotheses." *Philosophy of Science* 23(3): 237–46.

Jin, Y., Wang, L., Chen, G., Lin, X., Miao, W., & Fu, Z. (2014). Exposure of mice to atrazine and its metabolite diaminochlorotriazine elicits oxidative stress and endocrine disruption. *Environmental toxicology and pharmacology*, *37*(2), 782-790.

Kellert, S.H., H.E. Longino, and C.K. Waters (2006) Introduction: the pluralist stance, in S.H. Kellert, H.E. Longino, and C.K. Waters (eds.), *Scientific pluralism (Minnesota Studies in Philosophy of Science, Vol. 19)*, Minneapolis: University of Minnesota Press, vii– xxix.

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

Kloas, W., Lutz, I., Springer, T., Krueger, H., Wolf, J., Holden, L., & Hosmer, A. (2009). Does atrazine influence larval development and sexual differentiation in Xenopus laevis?. *Toxicological Sciences*, *107*(2), 376-384.

Kourany, Janet A. (2010). *Philosophy of Science After Feminism*. New York: Oxford University Press.

Krimsky, S. (2001). An epistemological inquiry into the endocrine disruptor thesis. *Annals of the New York Academy of Sciences*, *948*(1), 130-142.

Krimsky, S. (2002). *Hormonal chaos: the scientific and social origins of the environmental endocrine hypothesis*. JHU Press.

Kuhn, Thomas S. (1977) *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Lacey, H. (2005). *Is science value free?: Values and scientific understanding*. Psychology Press.

Lamb, J. C., Boffetta, P., Foster, W. G., Goodman, J. E., Hentz, K. L., Rhomberg, L. R., ... & Williams, A. L. (2014). Critical comments on the WHO-UNEP State of the Science of Endocrine Disrupting Chemicals–2012. *Regulatory Toxicology and Pharmacology*, *69*(1), 22-40.

Landis, W. G., and Yu, M. H. (2010). *Introduction to environmental toxicology: Impacts of chemicals upon ecological systems*. Taylor: Boca Raton, Florida.

Leonelli, S. (2009). On the locality of data and claims about phenomena. *Philosophy of Science*, *76*(5), 737-749.

Lewontin, R. C. (1974). *The genetic basis of evolutionary change* (Vol. 560). New York: Columbia University Press.

Longino, Helen E. (1995). Gender, Politics, and the Theoretical Virtues. *Synthese* 104(3): 383–97

Longino, Helen E.. (2001). *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

Longino, Helen E.. (2013) *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*. Chicago: University of Chicago Press.

Love, Alan (2008). Explaining evolutionary innovations and novelties: Criteria of explanatory adequacy and epistemological prerequisites. *Philosophy of Science*, 75: 874–886.

Ludwig, D. (2015). Ontological Choices and the Value-Free Ideal. *Erkenntnis*, 1-20.

Machery, E. 2009. *Doing Without Concepts*. (New York: Oxford University Press).

Margolis, E., & Laurence, S. (1999). *Concepts: core readings*. Mit Press.

Massimi, M. (2011). From data to phenomena: a Kantian stance. *Synthese*, *182*(1), 101-116.

Mayr, E. (1987). The ontological status of species: scientific progress and philosophical terminology. *Biology and Philosophy*, *2*(2), 145-166.

Mays, Vickie M., and Susan D. Cochran (2001) Mental Health Correlates of Perceived Discrimination Among Lesbian, Gay, and Bisexual Adults in the United States." *American Journal of Public Health* 91(11): 1869–76.

McLachlan, J. A., R. R. Newbold, and B. Bullock. (1975) Reproductive Tract Lesions in Male Mice Exposed Prenatally to Diethylstilbestrol." *Science* 190(4218): 991–2.

Mohai, P. (1994). Demographics of Dumping Revisited: Examining the Impact of Alternate Methodologies in Environmental Justice Research, The. *Va. Envtl. LJ*, *14*, 615.

Montazerhodjat, Vahid, and Andrew W. Lo. (2015) *Is the FDA Too Conservative or Too Aggressive?: A Bayesian Decision Analysis of Clinical Trial Design* (No. w21499). National Bureau of Economic Research.

Nagel, E. (1961), *The structure of science*. New York: Harcourt, Brace, and World.

Nelson, Arthur C.; Genereux, John; and Genereux, Michelle. (1992) Price Effects of Landfills on House Values. *Land Economics* 68 (4): 359–65.

Neurath, O. (1973). *Wissenschaftliche Weltauffassung: Der Wiener Kreis* (pp. 299-318). Springer.

Nisbet, Matthew C., and Chris Mooney. (2007). Framing Science. *Science* 316(5821): 56.

Nohynek, G. J., Borgert, C. J., Dietrich, D., & Rozman, K. K. (2013). Endocrine disruption: fact or urban legend?. *Toxicology letters*, *223*(3), 295-305.

O'Malley, M. A., Elliott, K. C., Haufe, C., & Burian, R. M. (2009). Philosophies of funding. *Cell*, *138*(4), 611-615.

O'Malley, M. (2007). Exploratory experimentation and scientific practice: Metagenomics and the proteorhodopsin case., *Hist. Phil. Life Sci*. 29(3).

Oreskes, N., & Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA.

Oudshoorn, Nelly. (2003). *Beyond the Natural Body: An Archaeology of Sex Hormones*. New York: Routledge.

Plutynski, A. (2005). Explanatory unification and the early synthesis. *The British journal for the philosophy of science*, *56*(3), 595-609.

Powers, Jack (2014) *Atrazine Research and Criteria of Characterizational Adequacy*. In: [2014] Philosophy of Science Assoc. 24th Biennial Mtg (Chicago, IL). *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*

Resnik, D. B., & Elliott, K. C. (2013). Taking financial relationships into account when assessing research. *Accountability in research*, *20*(3), 184-205.

Richardson, Sarah S. (2013) *Sex Itself: The Search for Male and Female in the Human Genome*. Chicago: University of Chicago Press.

Robert, J.S. (2004). *Embryology, epigenesis, and evolution: taking development seriously*. New York: Cambridge University Press.

Rohr, J.R., McCoy K.A. (2010a). A qualitative meta-analysis reveals consistent effects of atrazine on freshwater fish and amphibians", *Environ Health Persp* 118, 20–32

Rohr, J.R., McCoy K.A. (2010b). Preserving environmental health and scientific credibility: a practical guide to reducing conflicts of interest. *Conservation Letters* 3 143–150

Rooney, P. (1992, January). On values in science: Is the epistemic/non-epistemic distinction useful?. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (pp. 13-22). Philosophy of Science Association.

Rosenberg, A. (1997). Reductionism redux: computing the embryo. *Biology and Philosophy* 12:445–470.

Rothwell, E. (2010). Analyzing focus group data: content and interaction. *Journal for Specialists in Pediatric Nursing, 15*(2), 176-180.

Roughgarden, Joan (2009). *The Genial Gene: Deconstructing Darwinian Selfishness*. Berkeley, CA: University of California Press.

Rudner, Richard. (1953) The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science* 20(1): 1–6.

Saglio, P., & Trijasse, S. (1998). Behavioral responses to atrazine and diuron in goldfish. *Archives of Environmental Contamination and Toxicology*, *35*(3), 484-491.

Sankey, H. (2010). Scientific Method, in *The Routledge Companion to the Philosophy of Science*, Routledge.

Santos, T. G., & Martinez, C. B. (2012). Atrazine promotes biochemical changes and DNA damage in a Neotropical fish species. *Chemosphere*, *89*(9), 1118-1125.

Schaffner, K. F. (1969), The Watson-Crick model and reductionism, *British Journal for the Philosophy of Science,* 20, 325–348.

Schaffner, K. F. (1993). *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press.

Skelly, David K., Susan R. Bolden, and Kristin B. Dion. (2010) Intersex Frogs Concentrated in Suburban and Urban Landscapes." *EcoHealth*, *7*(3): 374–9.

Sober, E. (1999). The multiple realizability argument against reductionism. *Philosophy of Science* 66:542–564.

Solomon, K. R., Carr, J. A., Du Preez, L. H., Giesy, J. P., Kendall, R. J., Smith, E. E., & Van Der Kraak, G. J. (2008). Effects of atrazine on fish, amphibians, and aquatic reptiles: a critical review. *Critical reviews in toxicology*, *38*(9), 721-772.

Solomon, K. R., Giesy, J. P., LaPoint, T. W., Giddings, J. M., & Richards, R. P. (2013). Ecological risk assessment of atrazine in North American surface waters. *Environmental Toxicology and Chemistry*, *32*(1), 10-11.

Spike, Caroline A., Jason Bader, Valerie Reinke, and Susan Strome. (2008) DEPS-1 promotes P-granule assembly and RNA interference in *C. elegans* germ cells." *Development* 135(5): 983–93.

Spike, Caroline, Nicole Meyer, Erica Racen, April Orsborn, Jay Kirchner, Kathleen Kuznicki, Christopher Yee, et al. (2008) Genetic Analysis of the *Caenorhabditis elegans* GLH Family of P-Granule Proteins. *Genetics* 178(4): 1973–87.

Spirtes, P., Glymour, C., & Scheines, R. (2001). Causation, Prediction, and Search, Volume 1 of MIT Press Books.

Steel, D., & Whyte, K. P. (2012). Environmental justice, values, and scientific expertise. *Kennedy Institute of Ethics Journal*, *22*(2), 163-182.

Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science*, S65-S74.

Tabery, J.G. (2004). Synthesizing activities and interactions in the concept of a mechanism. *Philosophy of Science* 71:1–15.

Taylor, H., & Vickers, P. (2015). Conceptual fragmentation and the rise of eliminativism. *European Journal for Philosophy of Science*, 1-24.

Taylor, J. H. and Vickers, Peter (2015) Conceptual Fragmentation and the Rise of Eliminativism. [Preprint] URL: http://philsci-archive.pitt.edu/id/eprint/11320 (accessed 2016-04-08).

Thurman E M, Cromwell A E (2000). Atmospheric transport, deposition, and fate of triazine herbicides and their metabolites in pristine areas at Isle Royale National Park. *Environ Sci Technol* 34:3079–3085.

Van Der Kraak, G. J., Hosmer, A. J., Hanson, M. L., Kloas, W., & Solomon, K. R. (2014). Effects of atrazine in fish, amphibians, and reptiles: an analysis based on quantitative weight of evidence. *Critical reviews in toxicology*, *44*(sup5), 1-66.

Vom Saal, F. S., & Hughes, C. (2005). An extensive new literature concerning low-dose effects of bisphenol A shows the need for a new risk assessment. *Environmental health perspectives*, 926-933.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., ... & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, *5*(1), 14-26.

Waters, C.K (1990). Why the antireductionist consensus won't survive the case of classical Mendelian genetics. in A. Fine, M. Forbes and L. Wessels (eds.), *Proceedings of the biennial meeting of the Philosophy of Science Association*, Vol. 1, East Lansing, MI: Philosophy of Science Association, 125–139.

Waters, C.K (2007). The nature and context of exploratory research. *Hist. Phil. Life Sci.* 29(3).

Wells, Kentwood D. (1977). Territoriality and Male Mating Success in the Green Frog (*Rana clamitans*). *Ecology* 58(4): 750–62.

Wells, Kentwood D. (1978). Territoriality in the Green Frog (*Rana clamitans*): Vocalizations and Agonistic Behaviour. *Animal Behaviour* 26(4): 1051–63.

West-Eberhard, M. J. (1984). Sexual selection, competitive communication and species-specific signals in insects. *Insect communication*, 283-324.

Wilholt, Torsten. (2009). Bias and Values in Scientific Research. *Studies in History and Philosophy of Science Part A* 40(1): 92–101.

Wilson K, Hardy I. (2002). Statistical introduction of sex ratios- an introduction. In: Hardy ICW (ed) Sex ratios - concepts and research methods. Cambridge University Press, Cambridge, pp 48–92

Wilson, M. (2006). *Wandering Significance*. Oxford: OUP.

Wimsatt, W. C. (1976). Reductive explanation: A functional account. In *PSA 1974* (pp. 671-710). Springer Netherlands.

Wimsatt and Schank (2004). Generative entrenchment, modularity and evolvability: when genic selection meets the whole organism**.** in G. Schlosser and G Wagner, eds., *Modularity in Development and Evolution,* Chicago: U Chicago Press, 359-394.

Witschi, E. (1929). Rudimentary hermaphroditism and Y chromosome in *Rana temporaria*. *J. Exp. Zool*. 54:157–223.

Witschi, Emil. (1929) Development of Gonads and Transformation of Sex in the Frog. *The American Naturalist* 55(641): 529–38.

Zoeller, R. T., Bergman, Å., Becher, G., Bjerregaard, P., Bornman, R., Brandt, I., ... & Skakkebaek, N. E. (2014). A path forward in the debate over health impacts of endocrine disrupting chemicals. *Environmental Health*, *13*(1), 1.

Zuk, Marlene. (1993) Feminism and the Study of Animal Behavior. *BioScience* 43(11): 774–8.