Routledge Studies in Ethics and Moral Theory

RISK AND RESPONSIBILITY IN CONTEXT

Edited by
Adriana Placani and Stearns Broadhead

ROUTLEDGE

Risk and Responsibility in Context

This volume bridges contemporary philosophical conceptions of risk and responsibility and offers an extensive examination of the topic. It shows that risk and responsibility combine in ways that give rise to new philosophical questions and problems.

Philosophical interest in the relationship between risk and responsibility continues to rise, due in no small part to environmental crises, emerging technologies, legal developments, and new medical advances. Despite such interest, scholars are still working out how to conceive of the links between risk and responsibility, the implications that risks may have to conceptions of responsibility (and vice versa), as well as how such theorizing might play out in applied cases. With contributions from leading scholars, this volume brings together new work examining the interplay between risk and responsibility, exploring its varied philosophical aspects and applications to contemporary issues in law, bioethics, technology, and environmental ethics.

Risk and Responsibility in Context will be of interest to philosophers working in ethics, bioethics, philosophy of law, and philosophy of technology, as well as scholars and practitioners in law, health and science management, public policy, and environmental studies.

Adriana Placani is an appointed research fellow at the NOVA University of Lisbon's Institute of Philosophy (IFILNOVA).

Stearns Broadhead is a post-doctoral researcher at the University of Graz's Institute of Philosophy, working on the Austrian Science Fund-financed research project *Responsibility for Risks: Theory and Practice*.

Routledge Studies in Ethics and Moral Theory

Philosophical Perspectives on Moral Certainty
Edited by Cecilie Eriksen, Julia Hermann, Neil O'Hara, and Nigel Pleasants

The Guise of the Good
A Philosophical History
Francesco Orsi

The Making of the Good Person
Self-Help, Ethics and Philosophy
Nora Hämäläinen

Moral Teleology
A Theory of Progress
Hanno Sauer

Agent Relative Ethics
Steven J. Jensen

Moral Injury and the Humanities
Interdisciplinary Perspectives
Edited by Andrew I. Cohen and Kathryn McClymond

Experiments in Moral and Political Philosophy
Edited by Hugo Viciana, Antonio Gaitán, and Fernando Aguiar

Risk and Responsibility in Context
Edited by Adriana Placani and Stearns Broadhead

For more information about this series, please visit: https://www.routledge.com/Routledge-Studies-in-Ethics-and-Moral-Theory/book-series/SE0423

Risk and Responsibility in Context

Edited by Adriana Placani
and Stearns Broadhead

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2024
by Routledge
605 Third Avenue, New York, NY 10158

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Research results from: Austrian Science Fund (FWF): project number P-31527.

Published with the support of the Austrian Science Fund (FWF): PUB 1003-P.

FWF Austrian Science Fund

ISBN: 978-1-032-21912-7 (hbk)
ISBN: 978-1-032-23159-4 (pbk)
ISBN: 978-1-003-27602-9 (ebk)

DOI: 10.4324/9781003276029

Typeset in Sabon
by KnowledgeWorks Global Ltd.

Contents

PART III

PART IV

PART V

List of Contributors

Stearns Broadhead is a post-doctoral researcher at the University of Graz's Institute of Philosophy, working on the Austrian Science Fund-financed research project *Responsibility for Risks: Theory and Practice*.

Samantha Copeland is an assistant professor in the Ethics and Philosophy of Technology section of TU Delft's Department of Values, Innovation and Technology.

Christophe Depaus is a senior expert in safety strategy for the geological disposal of high-level radioactive waste at ONDRAF/NIRAS (Brussels).

Neelke Doorn is a distinguished Antoni van Leeuwenhoek Professor 'Ethics of Water Engineering' in the Ethics and Philosophy of Technology section of TU Delft's Department of Values, Innovation and Technology and Director of Education of the Faculty of Technology, Policy and Management.

RA Duff is an emeritus professor in Philosophy at the University of Stirling.

Jessica Nihlén Fahlquist is an associate professor of Practical Philosophy and a Senior Lecturer in Biomedical Ethics at Uppsala University's Centre for Research Ethics and Bioethics.

Benjamin Hale is an associate professor in the Philosophy Department and the Environmental Studies Program at the University of Colorado Boulder.

Sven Ove Hansson is a professor in the Department of Philosophy and History at the Royal Institute of Technology (Stockholm).

Madeleine Hayenhjelm is a senior lecturer in Philosophy at Umeå University.

Avram Hiller is an associate professor in the Philosophy Department at Portland State University.

Céline Kermisch is a lecturer at the École polytechnique de Bruxelles (Université Libre de Bruxelles) and a senior expert in ethics of science and technology.

Anne Ruth Mackor is a professor of Professional Ethics, in particular legal professions, at the University of Groningen's Faculty of Law, and she is chair of the North Sea Group, a research network that regularly organizes on- and offline seminars on the theory and methodology of evidence and legal proof.

Sven Nyholm is a professor of the Ethics of Artificial Intelligence at Ludwig Maximilian University of Munich, Germany.

Adriana Placani is an appointed research fellow at NOVA University of Lisbon, Institute of Philosophy (IFILNOVA).

Ibo van de Poel is Antoni van Leeuwenhoek Professor in Ethics and Technology at the School of Technology, Policy and Management at TU Delft.

Sabine Roeser is Antoni van Leeuwenhoek Professor in Ethics and Head of the Department of Values, Technology and Innovation, Faculty of Technology, Policy and Management at TU Delft.

Martin Sand is an assistant professor of Ethics and Philosophy of Technology at *TU Delft*.

Kenneth Shockley is a professor and Holmes Rolston Professor of Environmental Ethics and Philosophy at Colorado State University.

Steffen Steinert is an assistant professor at the Ethics and Philosophy of Technology section at TU Delft.

Acknowledgments

We thank the contributors to this volume for their efforts and time. We also thank Andrew Weckenmann at Routledge for his guidance and help, as well as three anonymous reviewers for their feedback. Harald Stelzer was crucial to our work, and we greatly appreciate his support.

The editors' work on this volume was conducted at the University of Graz's Institute of Philosophy as part of the research project *Responsibility for Risks: Theory and Practice* and at NOVA University of Lisbon's Institute of Philosophy (IFILNOVA). The editors gratefully acknowledge that their research was funded by the Austrian Science Fund (FWF): project number P-31527. Adriana Placani's work was also financed by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the Scientific Employment Stimulus - Individual Call - CEECIND/02135/2021.

1 Risk, Responsibility, and Their Relations

Adriana Placani and Stearns Broadhead

1.1 Introduction

Risk and responsibility are fertile topics of philosophical investigation. Often, but not always, they are considered separately. While responsibility has a long and varied history, risk as a topic of philosophical focus, in ethics at least, is not as longstanding (Erman and Möller 2018, 207; Hayenhjelm and Wolff 2011, E27). This volume examines risk and responsibility as continuous topics. That is, in its broadest formulation, the volume's contributions consider some of the many ways in which risk and responsibility relate to each other and combine in philosophy. Such combination, as this book shows, is not limited to a single account of risk or responsibility, nor is it applied to just one issue or within a lone context.

The contributions to this volume examine responsibility and risk by addressing issues that arise out of their interplay within various contexts, whether conceptual, legal, bioethical, technological, or environmental. As such, they raise new and challenging issues across a multitude of philosophical areas of investigation and, ultimately, scrutinize the complexities of the modern world through the lens of risk and responsibility. This points to why risk and responsibility merit such special attention: risk and responsibility, often but not always formulated as responsibility for risks, are at the heart of many central problems of the modern age. Moreover, discussions of how risk should be incorporated into moral and political theories, in which responsibility is a central concept but the concept of risk is less often addressed, are of central and growing interest in philosophy. Such academic interest is supplemented by the fact that risk management with its attendant responsibilities has become a topic of increased public concern (e.g., pandemic risks). Thus, a foray into the topic of risk and responsibility, examined in different contexts and applications, has now become crucial for understanding much of our present world and for guiding its future.

DOI: 10.4324/9781003276029-1

The spectra of topics and themes considered by contributors to this volume represent areas of research that continue to generate intense discussion. Part I problematizes the idea of control within both responsibility and risk conceptualizations. Part II addresses problems related to risk and responsibility that arise within the law in pre-trial detention and in the statistical use of probabilities in courts. Part III tackles considerations related to risk and responsibility in bioethics by examining luck egalitarianism, responsible risking, and public health risks. Part IV considers issues of risk and responsibility across the technological field by examining the role of emotions in the responsible innovation of risky technologies, artificial intelligence (AI), and radioactive waste management. Part V addresses the topic within environmental ethics by examining a host of considerations pertaining to individual climate risks and resilience.

Regardless of the perspective adopted on the topic at hand, exploring the relationships between responsibility and risk requires clear notions of each. This introduction focuses on doing just that, analyzing risk and responsibility separately and allowing their synthesis and application to come out primarily in the volume's constitutive chapters. This introduction does, however, identify the concepts and topics explored and elaborated in the respective chapters. In this way, the introduction helps to contextualize and explain the concepts of risk and responsibility, and it also helps to make sense of the relationships between them as discussed in the rest of the volume.

The rest of this introduction starts with an examination of the concept of risk, detailing some of its definitions, dimensions, and conceptualizations. This is followed by an exploration of the concept of responsibility, which outlines some of its senses and dimensions. As noted, these sections are not meant to be exhaustive treatments but rather introductions to the topic by way of outlining its constituent parts. Finally, an overview of each of this volume's contributions highlights the ways that this volume brings together the concepts of risk and responsibility.

1.2 Risk

Philosophical interest in risk has been intensifying. This is understandable, in part, because of the pervasiveness of risks. Consider, for example, that most of our decisions are made under conditions of risk or uncertainty about the possible consequences of our actions or omissions (e.g., what career to choose, whether to cross the street). Moreover, growing interest in the topic seems also attributable, at least partly, to relatively new concerns, such as the risks posed by anthropogenic climate change, new, emerging, and future technologies, as well as the use of preemptive legal measures. This diversity of interests has meant that, similar to responsibility (as we will see), risk has accrued many senses (Bradbury 1989; Hansson 2004; Renn 1992; Shrader-Frechette 1991; Thompson and Dean 1996).

1.2.1 Definitions

There is not one definition of risk; there are many. This is because the concept of risk is used in a variety of disciplines with different specialized meanings but also in everyday life where the meaning of risk tends to be much looser (i.e., non-specialized). Both the specialized and non-specialized understandings of risk are important, and any precedence of one over the other seems to depend on, *inter alia*, one's context and aims.[1] The following section will provide an overview of some definitions of risk as well as detail dimensions of risk that can illuminate the concept.

In everyday life, in non-specialized contexts, risk usually refers to something undesirable that is possible but not certain to occur. For example, a parent might tell their child to wash their hands after playing outside because they might get sick otherwise. The risk in the example is the risk of sickness, which is an undesirable outcome that may or may not occur.

In specialized or technical domains, risk admits of many perspectives depending on the area of investigation (e.g., psychology, economics, engineering, sociology, philosophy).[2] Categorizing risk conceptions across various disciplines is a challenging task; however, Hansson (2004, 10) provides a list of conceptions of risk that is a useful guide to the more prominent uses of the term:

1. Risk as an *unwanted event* that may or may not occur
2. Risk as the *cause* of an unwanted event that may or may not occur
3. Risk as the *probability* of an unwanted event that may or may not occur
4. Risk as the fact that a decision is made under conditions of *known probabilities* (this is a decision under risk, which is usually contrasted with a decision under uncertainty)
5. Risk as the statistical *expectation value of unwanted events* that may or may not occur

Hansson's example of the risks associated with cigarette smoking can help clarify the differences in the meanings of risk mentioned above. Consider, then, that:

> Lung cancer is one of the major risks (1) that affect smokers. Smoking also causes other diseases, and it is by far the most important health risk (2) in industrialized countries. There is evidence that the risk (3) of having one's life shortened by smoking is as high as 50%.

> The total risk (4) from smoking is higher than that from any other cause that has been analyzed by risk analysts. The probabilities of various smoking-related diseases are so well-known that a decision whether or not to smoke can be classified as a decision under risk (5).
>
> (Hansson 2004, 10)

Hansson's list of risk conceptions is not exhaustive. Moreover, there is no consensus over either the definition or conception of risk that is best suited for philosophical investigation. Some philosophers (Buchak 2014; Pritchard 2015; Shrader-Frechette 1991) criticize the specialized definitions as being too narrow, and the non-specialized everyday sense of the term risk is sometimes favored.

Despite differences, all the definitions have in common the fact that they regard risk either to be or involve something undesirable about which there is some lack of knowledge regarding its occurrence. With regard to knowledge, the first two definitions of risk are non-probabilistic, while the last three refer to probabilities. In other words, we can say that the last three, as opposed to the first two definitions, seek to quantify the degree of knowledge an individual has over the possible occurrence of the unwanted event by employing the notion of probability. The question then becomes what probability is.

1.2.2 Probability

Probability interpretations seem to be no less abundant than those of risk. The complexities of understanding probability are so challenging, in fact, as to prompt Bertrand Russell to state: "Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means" (Stevens 1951, 44). Interpretations of probability are varied (e.g., classical, logical/evidential, subjective, frequentist). However, it bears noting that all probability interpretations agree on at least the fact that probabilities are numbers between zero and one that can attach to certain types of propositions and subjected to a probability calculus (e.g., Kolmogorov's Probability Calculus, Bayes' Theorem).[3] If an event cannot occur, then its probability is zero; while if an event is certain to occur, its probability is one.

In line with the current philosophical literature on the topic, the following will focus on the objective and subjective interpretations of probability, which give us the objective and subjective understandings of risk. Briefly stated, the subjective interpretation views risk and probability as, fundamentally, a matter of some kind of belief; the objective interpretation regards them as features of the world that exist, in a relevant sense, independently of human belief.

To the former, the subjective Bayesian theory sees the probability of a proposition as someone's degree of belief, credence, or confidence in that proposition.[4] For example, for Jane, the probability that it will rain tomorrow is her degree of belief that it will rain. If her degree of belief is 1/3, then her degree of belief that it will not rain is 2/3; her degree of belief that it will either rain or not rain is 1. We might say, then, that probabilities in the Bayesian theory are numerical measures of particular individuals' confidence in some proposition(s). These measures can be arrived at in different ways depending on one's version of the theory (e.g., on the basis of agents' betting behaviors).[5]

Some versions of the subjective theory posit that objective probabilities (e.g., frequencies, propensities) do not exist and that the only measure of probability is individualistic. One difficulty with such a view is that probability judgments can vary widely from person to person, especially in the absence of constraints on what ought to count as a rational belief. Consider Gillies' version of the subjective theory in order to see this:

> Probability is [...] defined as the degree of belief of a particular individual, so that we should really not speak of the probability, but rather Ms. A's probability, Mr. B's probability or Master C's probability.
>
> (Gillies 2000, 53)

Such an interpretation of probability would give rise to as many estimations of risk as there are beliefs with potentially no way of privileging one over the other (Oberdiek 2017, 22). Such an extreme version of a personalist account can be tempered by constraints of what counts as a rational belief, by coherence demands or by referring to the evidence for one's beliefs. It should be noted, however, that the latter would still depend on one's subjective evaluation of the support that the evidence provides to one's beliefs (Oberdiek 2017, 22). Nevertheless, versions of the Bayesian theory that posit, *inter alia*, norms requiring that degrees of belief respect the axioms of probability, empirical norms that require an agent's degrees of belief to be calibrated with her evidence, and logical norms that require degrees of belief underdetermined by evidence to be as equivocal as possible can rein in the pitfalls of subjectivity.

In turn, objective accounts of risks typically rely on frequentist interpretations of probability.[6] Frequentists view a risk of an event, E, as the frequency with which E occurs in the general population or some other reference class that is selected. The frequency with which the risk manifests in the reference class is taken to be an objective and scientifically verifiable fact. For example, the probability that a man over the age of 60 is diabetic equals the proportion of men over the age of 60 (the reference class) who are diabetic.

In spite of differences between versions of the frequentist theory, probability can be defined in accordance with a frequentist view as relative to a reference class that must be general, infinite, or at least very large (Gillies 2000, 88–112). Thus, probability in this theory is conceived in general terms. Reference is made to general attributes (e.g., diabetes) and general reference classes (e.g., men over 60), which raises the reference class problem and the problem of the single case. The former appears because probability for the frequentist is, basically, the long-run frequency of repeatable experiments. For example, the probability that a fair coin will land tails is 0.5 because were we to flip the coin enough times, we would get tails 50% of the time. However, singular, unique events are not repeatable by definition. To capture the problem of the single case for the frequentist theory, consider Reichenbach's take on it:

> I regard the statement about the probability of the single case, not as having a meaning of its own, but as representing an elliptical mode of speech. In order to acquire meaning, the statement must be translated into a statement about a frequency in a sequence of repeated occurrences. The statement concerning the probability of the single case thus is given a fictitious meaning, constructed by a transfer of meaning from the general to the particular case.
>
> (1949, 376–7)

Then, we also have the problem of the reference class. This is associated with the fact that the probability of an event occurring can change depending on how it is classified, and the same event can be classified in a variety of ways on the basis of it belonging in different reference classes.[7] The reference class problem appears in Venn (1876, 194), where he writes: "It is obvious that every individual thing or event has an indefinite number of properties or attributes observable in it, and might therefore be considered as belonging to an indefinite number of different classes of things." There is yet no established solution to the above problems for the frequency theory (Oberdiek 2017, 27).

1.2.3 Risk and Uncertainty

Having outlined some of the main interpretations of probability and some of their problems, we can return to the "lack of knowledge" aspect that was present in all of the definitions of risk stated earlier. To this, it bears to note that, in decision theory, lack of knowledge is categorized into two main types: risk and uncertainty. Thus, a distinction between risk and uncertainty is drawn, and this can be said to follow the probabilistic/non-probabilistic divide mentioned before. Knight draws the distinction in the following way:

> To preserve the distinction [...] between the measurable uncertainty and an unmeasurable one we may use the term "risk" to designate the former and the term "uncertainty" for the latter.
>
> (Knight 1921, 233)

In decision-making under risk, possible outcomes and their probabilities are known; while in decision-making under uncertainty, probabilities are either not known at all or known with insufficient precision (Hansson 2004, 11). It is not often that probabilities are known with certainty; however, when data is available, it becomes possible to determine probabilities that can be called objective in the frequentist sense (Möller 2012, 63). Often, frequency data will have to be supplemented or perhaps even supplanted by expert judgment (Ibid, 63). Such expert judgments cannot be construed as objective fact, but it is not merely subjective (in the classical sense) either because, as noted before, subjective probabilities measure a person's degree of belief that may need to satisfy various norms (e.g., probability axioms, logical norms) but need not correlate with objective frequencies (Ibid). Thus, such expert judgments might be better described as subjective estimates of objective probabilities (Ibid).

Knightian uncertainty refers to cases where we lack probabilities, but it should be noted that some theories seek to measure uncertainty by reducing it to probability.[8] For example, for subjectivist theories, probability represents all aspects of a decision-maker's lack of knowledge (Ibid, 65). Bayesians conceive all rational decisions as admitting of probabilities because for them rational decision-makers always assign a probability value to each potential outcome be it implicitly or explicitly (Ibid). Faced with new information, agents may also change their probability assessments (in accordance with Bayes' theorem), but they always assign determinable probabilities to all states of affairs (Ibid).[9]

1.2.4 Disvalue

The above sections sketched out some of the ways in which we might come to understand the lack of knowledge aspect that is present in all of the definitions of risk provided at the outset. There is yet one other aspect that is common to most interpretations of risk, be it in an implicit or explicit manner; namely, risk refers to something negative, unwanted, which is to be avoided. It is actually this feature of risk that most easily connects to responsibility questions in virtue of its normativity. If something should be avoided, then questions about whose responsibility it is to avoid it (prospective responsibility sense) and questions about whose responsibility it is if it is not avoided (retrospective responsibility sense) can arise.

In order to avoid charges of triviality, the negative dimension of risk is usually construed as some kind of harm. This construal has the added advantage of providing a unique currency that can be measured and compared. Nevertheless, this is easier said than done. The concept of harm might be basic, but it is not devoid of controversy when it comes to its conceptualization. Harm can be defined as a setback to interests, but this popular conception is not without problems. For example, some critics have pointed out that defining harm in terms of interests is defining one unclear concept in terms of another unclear concept (Miller 2010, 119). Furthermore, although some cases of harmfulness might be easily compared (e.g., losing two arms is worse than losing one, a severed spine is more serious than a headache), many harms are exceedingly difficult to measure and/or compare (e.g., how do we measure psychological harms, how many headaches equal a severed spine, how do we conceive of the harm of a species becoming extinct). Even when you restrict harm to just one kind – the harm of death – questions remain, such as: Is death a harm (not for the Epicurian)? Is the harmfulness of an 80-year-old's death the same as that of 19-year-old's? Is the harmfulness of the death of a cancer patient who is in severe chronic pain the same as that of a healthy 21-year-old? Measurements that consider both the quantity and the quality of life, such as HALY (health-adjusted life years) with its types, QALY (quality-adjusted life years), and DALY (disability-adjusted life years), are controversial and measuring and comparing the severity of harms remains a contested area.

1.2.5 Multidimensional Conceptions

The two dimensions of risk considered above, lack of knowledge and adverse consequences, can be found in many risk conceptions and are constitutive of the most common definition in risk analysis, which equates risk to the expected value of unwanted events. However, there are other interpretations of risk, which introduce new dimensions. Typically, conceptions of risk in psychology, social science, and moral theory are sensitive to contextual factors as well, and they include different aspects within their risk conceptualizations besides the two that were mentioned, such as who runs the risk, whether the risk is imposed or voluntarily incurred, or whether the risk is natural or human-made (van de Poel and Fahlquist 2012, 881). The following will highlight some of these conceptions, albeit in brief, in order to exemplify the rich diversity of views on the topic.

The psychological literature on risk, which has been developing since at least the 1960s, has shown that lay people include a variety of elements in how they perceive and understand risks (Slovic 2000). Such elements include fear, perceived benefits, time delays, voluntariness, familiarity, controllability, catastrophic potential, and exposure (van de Poel and Fahlquist 2012, 881). Studies have found large differences between so-called real or objective risk and perceived risk in some cases. To exemplify, Lichtenstein et al. (1978) found that bad outcomes that were easier to recall tended to be thought of as more likely to occur. This phenomenon is described as the availability heuristic (Tversky and Kahneman 1974). Risk assessment is also driven by affective states. Lay people tend to exploit the so-called affect heuristic, which refers to the fact that people make judgments based on representations of objects or events that are marked with valenced affect (Slovic et al. 2002). Sometimes the fact that lay people employ different views and estimates of risk than experts is seen as a sign of their irrationality (van de Poel and Fahlquist 2012, 881). However, this interpretation presumes that the technical conception of risk is the right one and that lay people should be educated to comply with it and that, were they to be so educated, they would come to understand risk in line with its objective understanding (van de Poel and Fahlquist 2012, 881). However, many authors have argued that the elements included by lay people in their risk conceptualizations are relevant (e.g., for risk acceptability and management) and provide richer and sometimes more appropriate conceptions of risk than those of the experts (e.g., Roeser 2006, 2007; Slovic 2000).

Other richer conceptions of risk come from the literature on risk ethics. Traditionally, ethical theories have dealt with assessing moral problems in contexts of certainty where actions were assumed to have determined outcomes. This may have been because of the benefits of keeping at least some things relatively free from complicating factors, as well as the assumption of a division of labor between ethics and decision theory. Contending with problems associated with a lack of knowledge was seen as a task belonging to decision theory (Hansson 2003, 291).

Moreover, the major ethical branches, such as deontology, utilitarianism, and contractualism, suffer from particular weakness when it comes to considering risk within their respective frameworks due in part to their theoretical commitments and to the fact these were adopted with the implicit assumption of certainty vis-à-vis outcomes.[10] Still, contemporary philosophers have recognized that risks are pervasive, introduce genuine ethical dilemmas, and so they must be dealt with in spite of the many challenges they bring.

The diversity of ethical views on the topic is great, but with regard to assessing the moral acceptability of risks, many agree that this depends on more factors than those allowed by the standard technical definitions (i.e., the probability of harm combined with the severity of harm). Relevant factors include voluntariness, justice considerations, rights-based considerations, risks/benefits distribution, responsibilities (e.g., role responsibility), justifiability, and the availability of alternatives (Asveld and Roeser 2009; Caney 2009; Hansson 2009; Kermisch 2012; Shrader-Frechette 1991; Thomson 1986; van de Poel et al. 2012). Interestingly, Jonathan Wolff (2006) develops a new model of the anatomy of risk, which integrates several other factors, including responsibility-related ones, in the definition of risk besides probability and magnitude of harm. Wolff argues that to provide an adequate account of the factors that must be included in order to decide how to manage particular risks, attention must be given to cause, hazard, probability, fear, blame, and shame. In this way, Wolff explicitly connects factors from public perception and responsibility within the definition of risk itself. With regard to the latter, different human-made risks may be different in acceptability depending on whether they were caused by culpable or non-culpable behavior and on the type of culpable behavior (e.g., malice, recklessness, negligence, or incompetence) that caused them.

In the social sciences, we find many conceptions of risk, but perhaps the most influential has been that of Ulrich Beck (1992). In *Risk Society*, Beck writes about the many definitional struggles surrounding risks (e.g., over their scope and scale, degree, and urgency), the multitude of definitions themselves, as well as the agglomeration of misunderstandings and antagonisms given these issues (Placani 2017). Beck advances his own understanding by writing that: “Risk may be defined as a systematic way of dealing with hazards and insecurities induced and introduced by modernization itself” (Beck 1992, 21). The definition he provides ties the concept of risk to that of reflexive modernization, which is Beck’s call to confront and reflect upon the uncertainties of the modern age (Placani 2017). Beck sees risk as a social construct, which is historically a recent phenomenon that is closely tied to the idea that risks depend on decisions (Beck 1992, p. 183). In Beck’s risk society, all hazards are seen as depending on human choice and, hence, are, according to his definition, conceived as risks. As a result of this, in contrast to the industrial society, the principal issue in the risk society concerns the allocation of risk rather than that of wealth (van de Poel and Fahlquist 2012, 881–2).

1.2.6 Concluding Risk

The above section illustrated some of the more prominent conceptions of risk, but there are yet others. Whether richer senses of risk or sparser ones should be preferred remains an open question whose answer will likely depend on things such as context, aims, preferences, and theoretical commitments. Perhaps among the many differences in risk conceptions, commonalities were adumbrated as well. Still, the concept of risk remains contested and its dimensions are not yet fully explored. Nevertheless, the above concepts should serve the reader well because all the contributions to this volume, to varying extents, rely, as well as expand, on the understandings of risk discussed. However, before we proceed to discussing the entries to this volume, we need to explicate one more piece from our theoretical puzzle – the concept of responsibility.

1.3 Responsibility

The concept of responsibility is complex and has a variety of senses as well as uses, such as liability, role, capacity, and causal responsibility (Boxer 2014; Hart 1968). Given this complexity, it becomes necessary to provide an overview of some of the more prominent responsibility senses that can illuminate the concept and that are also featured in the contributions to this book. This will provide initial clarity to the matter and lay out a conceptual framework that will carry forward.

The place to start for most discussions of responsibility is with HLA Hart's taxonomy of responsibility's various senses (1968, 211–30). Hart distinguishes between four main conceptions of responsibility, and, as these are foundational, the following will briefly outline them. However, Hart's senses do not exhaust the concept, and the following will add to the explication a separate exploration of moral responsibility, temporal views on responsibility (i.e., its backward- and forward-looking senses), as well as collective responsibility. Admittedly, this still leaves out other senses of responsibility. The diversity of views on this topic is simply too great to be captured here.

As noted, Hart identifies four types of responsibility. A slim overview of the taxonomy is (1) causal responsibility, (2) capacity responsibility, (3) role responsibility, and (4) legal liability responsibility with subtypes (Cane 2002, 29).

1.3.1 Causal Responsibility

This first sense of responsibility, causal responsibility, is concerned with identifying agents or entities that bring about events or states of affairs. Consider that "Julia broke the window" is a way of saying "Julia is responsible for the window breaking." Causal responsibility assigns an agent or entity as cause of an event or state of affairs based on that agent or entity's involvement in it. Such causal involvement could be all-or-nothing (e.g., based on counterfactuals) or admit of degrees. Whatever kind of causal involvement is at stake it is not sufficient for inferring moral responsibility from it. This point highlights the way in which causal responsibility is neither a substitute nor a marker for moral responsibility even if the two may sometimes coincide in a single circumstance.

First, being a cause of an event or state of affairs is not sufficient for moral responsibility. Consider a slightly more detailed scenario: "Julia accidentally (without foreseeing, knowing, or intending) breaks a window." Julia caused the break, yes, but Julia accidentally broke the window because she got pushed by James. To be a cause in this respect (admittedly bracketing any further complications) is not sufficient grounds to be judged morally blame- or praise-worthy, good or bad, which are typical of judgments of moral responsibility. Julia's moral powers were not at work in breaking the window.

Second, certain causally responsible entities lack relevant moral agency for moral responsibility. Consider another scenario: "The tree branch broke the window." Yes, the tree branch caused something – a broken window – but to identify the tree branch as a moral agent would be absurd according to current science about trees. There are not, in other words, sufficient grounds for moral evaluation of the tree in ways that might apply to agents in similar circumstances.

1.3.2 Capacity Responsibility

Hart identifies capacity-responsibility in the following way. The expression "he is responsible for his actions is used to assert that a person has certain normal capacities [...] those of understanding, reasoning, and control of conduct: the ability to understand what conduct legal rules or morality require, to deliberate and reach decisions concerning these requirements, and to conform to decisions when made" (Hart 1968, 227). This list is not exhaustive, but it identifies capacities for rational agency.

The powers of reasoning and understanding are among the rational capacities of agents, while the capacity to control their conduct enables them to express these rational capacities in action (Raz 2010, 4). In other words, people are responsible for their conduct because they are rational agents, and as rational agents (Ibid). However, people are not responsible in this way if they lack capacity-responsibility or if the powers of rational agency constituting it are temporarily suspended or disabled (e.g., when people are asleep, under deep hypnosis, or when sensory deprivation is such that they cannot use their rational capacities) (Ibid). No doubt more can be said about capacity responsibility, but the topic will be addressed further in the discussion on moral responsibility.

1.3.3 Role Responsibility

Role responsibility describes an agent's responsibility due to her being charged with a duty or obligation to achieve or contribute to the accomplishment of a state of affairs. For example, "the supermarket's night manager must ensure that all newly received grocery products are stocked on the shelves." The agent has certain role-dependent duties ("responsibilities"). (There is another usage of role responsibility, which we set aside after mentioning it, as the comportment of an agent within the role they have. For example, "she is a responsible night manager who makes sure that all groceries are stocked before her shift ends." The descriptor here highlights that the agent takes the role's requirements seriously).

Whereas causal responsibility describes a relationship between (past) conduct causing an outcome (e.g., "she broke the window"), role responsibility involves a future-oriented or prospective responsibility to fulfill a prescribed duty or obligation associated with the role itself. One may fail or succeed in meeting her role-dependent obligations; however, the basis of responsibility is having assumed the role (with its duties or obligations). This, in effect, means one is on the hook in virtue of occupying a role, not necessarily having caused an outcome. Some careless stock clerk, not the supermarket's night manager, may have dropped all the palettes of groceries, thereby nixing them from being stocked on the shelves, but the night manager may nevertheless be (ultimately) responsible because she is the manager.

1.3.4 Legal Liability Responsibility with Subtypes

Legal liability responsibility refers to responsibility-based conditions of legal liability – for instance, to pay compensation, fines, restitution, or to be imprisoned. Broadly, when legal rules require one to act or abstain from action, one who breaks the law is usually liable, according to other legal rules, to some form of punishment (Hart 1968, 215). Punishment may be issued not only for one's own offences, but also for those of others. For example, an employer may be liable and suffer punishment for some offence committed by their employee. In law, such vicarious responsibility is a form of "strict" responsibility, which is responsibility regardless of fault (Cane 2002, 39).

In the legal context, it is also worth mentioning, given the interests of this volume, that negligence law allocates risk of liability for damages by holding people responsible for negligently bringing about certain harms (Raz 2010, 7). Negligence can be defined as a failure to meet a standard of behavior or a level of care that is established by law for the protection of others against unreasonable risk of harm. Key factors that help determine whether an action falls short of reasonable care are the foreseeable likelihood that the action will result in harm, the foreseeable gravity of that harm, and the burden of safeguarding against the risk of harm.

Moral liability responsibility is analogous to legal liability responsibility and thus may be considered a subtype. The differences between the two reside in the conditions for incurring each, respectively. According to Hart, we may define moral responsibility by substituting "liable to punishment" with "deserving blame" or "blameworthiness" and substituting "liable to be made to pay compensation" with "morally bound to make amends or pay compensation" (Hart 1968, 225). According to Hart, such responsibility depends on certain conditions that are related to the character or extent of a person's control over their own conduct, or to the causal or other connections between the person's action and harmful occurrences, or to his relationship with the person who actually did the harm (Hart 1968, 225). Given its prominence in the philosophical literature, it is more profitable to consider moral responsibility along with some of its conditions separately.

1.3.5 Moral Responsibility

The analysis of causal, role, capacity, and legal liability responsibility sheds light on aspects of moral responsibility. As discussed regarding causal responsibility, being a cause is not sufficient for moral responsibility, and some causally effective entities lack the requisite capacity for moral responsibility. With respect to role responsibility, although an agent may be morally responsible for failing to fulfill a duty, it is not a necessary condition of moral responsibility (Zimmerman 2016, 249). A night manager of the supermarket can be (role) responsible for a night stock clerk's failures, and she might also be morally responsible because she purposely acted to cause them. This latter conclusion is contingent on other facts and conditions. As for legal liability responsibility, as before, Hart advances a conception of moral responsibility understood as liability to blame or praise.

Nevertheless, merely asking what moral responsibility is belies the many disputes on a range of topics within subsets of the moral responsibility literature.[11] The moral responsibility literature largely thwarts attempts to describe moral responsibility using unitary, universally accepted definition (Buckareff, Moya, and Rosell 2015, 2; Fischer and Ravizza 1998, 10). Even so, there is some basic consensus and the following abstracts from various accounts to distill and examine two oft-noted conditions – the epistemic and control conditions (Ginet 2007; Haji 1998; McKenna 2008; Mele 1995; Pereboom 2014).

When speaking about moral responsibility, as foreshadowed above, something more is needed for praise or blame; namely, a capacity to behave as an agent able to be responsible (call it responsible agency). Responsible agency consists of two conditions. One is an epistemic condition and the other is a control condition, both of which are considered individually necessary and jointly sufficient for an agent's moral responsibility. Roughly speaking, the epistemic condition concerns an agent's cognitive state when acting, and the control condition refers to an agent's control (or freedom) in acting. Moral responsibility (disputably) requires an agent to satisfy these conditions.

The epistemic condition is complex in that it consists of capacities and qualities that an agent has. That is, arguably, only an agent who satisfies these conditions is a responsible agent who can be morally responsible (Wieland 2017). In the words of Oshana (2015), the epistemic condition entails that:

> The responsible agent is self-aware, that they are rational, that they are not ignorant of the circumstances in which they act, that they are cognizant of and able to act within established moral guidelines, and that they are responsive to reasons to adjust or amend their behavior in light of these guidelines. In order to be held responsible, the moral agent must know that doing a particular act (or an act of a given type) or cultivating a particular trait of character (say, jealousy, rage, bigotry) is right or is wrong. The moral agent may be held responsible if, suffering from none of the conditions that exempt a person from responsible agency, they should have known the nature of the act or trait, and could have been motivated by the relevant moral guidelines.
>
> (Oshana 2015, 13281)

In summary, the epistemic condition involves a complex cognitive capacity or awareness that marks agents as appropriate subjects of moral responsibility (i.e., candidates for blame). Responsible agents (to be designated as such) must have a requisite mental capacity (or acted or brought about a state of affairs in light of this capacity) in acting or (disputably) omitting to act (Talbert 2016, Ch. 5).

The aspect of capacity described by the epistemic condition identifies a necessary feature of responsible agency, but it does not suffice (under the standard meaning). The further challenge is to show how capacity responsibility (responsible agency) is or may be related to moral responsibility. After all, that an agent is a responsible agent in the relevant sense does not necessarily entail her moral responsibility for an action or resultant state of affairs. As Christopher Kutz describes it, "being responsible, in this sense, simply is a matter of having the competency of self-government" (2012, 549).

The control condition specifies the type or degree of control with which an agent acts (Talbert 2016, Ch. 1). An admittedly rough, but still viable, description of the requisite degree of control is voluntariness – traditionally construed as an agent being able to do otherwise (Corlett 2006; Frankfurt 1969, 11). This condition highlights that the sorts of actions or states of affairs relevant to responsible agency and moral responsibility are ones that agents accomplish through their own guidance or authorship, not coercion or manipulation, and so on (Fischer and Ravizza 1998, 12). Without this control over behavior, then attributions of blame or praise would incorrectly and perhaps unjustly target agents.

It bears noting that the bulk of the philosophical literature on responsibility has focused on moral responsibility understood as the blameworthiness of agents (e.g., Fischer and Ravizza 1998; Wallace 1994). In order to establish blameworthiness, a number of conditions have been put forward, among which we find those already identified, such as moral agency (capacity sense), causality (causal sense), freedom (control condition), and knowledge (epistemic condition). In addition, a condition that was implied, but not explicitly acknowledged, is that of wrongdoing. With regard to this, in order to elicit a justified attribution of blame, an agent must have done something wrong.

The conditions mentioned are common to many otherwise contrasting philosophical accounts, even though the relative weight and formulation that they are assigned may differ (Braham and Van Hees 2012; Cane 2002; Fischer and Ravizza 1998; Wallace 1994; Zimmerman 1988). This is not to say that they coexist in perfect harmony across accounts. Disagreements exist and tend to take one of the two forms: (1) the precise content of each of the conditions and (2) the necessary and/or jointly sufficient status of the conditions. Nevertheless, the responsibility conditions identified match well with commonsense morality and the specialized literature.

1.3.6 Collective Responsibility

The picture of responsibility that emerges from the above may seem to have an individualistic bent inasmuch as it asks questions about responsibility at an individual level. However, such questions may well be asked at a collective level as well. Thus, there are conceptions that seek to account for responsibility by focusing not just on individuals, but also on collectives.

Contra individualistic accounts of responsibility, collective responsibility accounts, do not restrict responsibility (e.g., causal, blameworthiness, role, liability) to individual agents. Instead, they focus on groups or collectives. Responsibility in such cases may be traced to collective intentions and collective actions taken by groups *qua* groups and distinct from individual members of such groups. However, much debate surrounds the moral agency of groups in general and the possibility of group intentions in particular, the distribution of collective responsibility to individual members, and the attribution of collective responsibility in particular cases (e.g., climate change, wars). It is unlikely that the debates surrounding the very possibility and content of collective responsibility will be settled any time soon (Copp 2007; French 1984; Gilbert 1989; Miller 2010; Narveson 2002).

In fact, there are philosophers who argue that collective responsibility does not exist, as individuals are the sole bearers of responsibility (e.g., Lewis 1948). However, it seems imperative to recognize that there are harms that cannot be traced to individuals acting alone and must be understood in their collective dimension. Considering the impact of multinational corporations, such as banks and oil producers, that behave badly and cause harms (e.g., environmental, economic), it seems crucial to find ways to hold such actors accountable for their actions (van de Poel and Fahlquist 2012, 892). It also seems crucial to recognize that, sometimes, collective actors (e.g., states, multinationals) are in a unique position to address such impacts, at least, for pragmatic reasons related to their capacity to do so.

Moreover, it seems plain to say that many risks in society admit of responsibility perspectives that are at the same time a matter for the individual and the collective (e.g., climate change risks, health risks, traffic-related risks). Consider climate-change-related risks. It seems plausible to argue that both individuals and governments have a responsibility to address these. Collectivist accounts (Cripps 2013; Sinnott-Armstrong 2005) hold that individuals' unilateral attempts to curb emissions are futile because the individual cannot make a difference and what is needed is large-scale collective action, which is the responsibility of states and national governments. Although virtually no scholar would deny the need for coordinated state action, it also seems plain to say that if it were truly the case that individual carbon dioxide emissions made no difference *at all*, then anthropogenic climate change would not occur (Hiller 2011). Moreover, that individuals have a responsibility to reduce their carbon dioxide emissions is a powerful intuition shared by many and argued forcefully (Baatz 2014; Berkey 2014; Broome 2012; Fruh and Hedahl 2013; Raterman 2012). Finally, it seems undeniable that action is needed at both individual and collective levels at least because individuals have a role to play in securing governmental action that is crucial for curbing emissions, as well as holding their governments responsible for such action or inaction.

1.3.7 Backward and Forward-Looking Responsibility

Crosscutting the senses of responsibility mentioned above is a view of responsibility that understands the concept in terms of its temporal dimensions. That is, another way of looking at responsibility (individual or collective) is by considering it in its backward- and forward-looking senses.

As mentioned before, most of the philosophical literature has focused on moral responsibility understood as blameworthiness. This is a backward-looking responsibility sense that looks to the past in order to see who, when, and under what conditions an agent is blameworthy, should be held to blame or be blamed for some action or outcome. It should be noted that the causal and liability senses of responsibility discussed above are also primarily backward-looking as they typically refer to something that has already occurred. However, the liability sense also admits of a forward-looking dimension to the extent that an agent is supposed to do something in the future in order to account for her actions, pay compensation, fines, restitution, etc.

The traditional focus on backward-looking moral responsibility understood as blameworthiness is complemented by forward-looking notions of responsibility, which typically address responsibility either on consequentialist grounds (e.g., Goodin 1995) or by relying on virtue or care ethics (e.g., Ladd 1991; Williams 2008). As related to the preceding analysis, the role responsibility sense discussed is forward-looking because it relates to responsibility as the obligation or duty to see to it that something is or will be the case. Responsibility as virtue is also understood as forward-looking (e.g., Bovens 1998; Ladd 1991) as it relates to responsibilities an agent assumes for herself and to certain attitudes or character traits she ought to cultivate.

When it comes to risks, the two temporal dimensions of responsibility offer clear connections. The clearest way in which backward-looking responsibility relates to risks is once they have materialized. In such cases, questions such as the following become pertinent: Who is responsible for the manifestation of the risk? Who should be held blameworthy for the risk manifesting? Who should compensate for the risk? The clearest way in which forward-looking responsibility relates to risks regards their management and prevention. In such cases, questions such as the following become pertinent: Who is responsible for averting the risk? Who is responsible for minimizing the risk and its impact? Who should be held liable for compensation if the risk manifests? These questions do not exhaust the queries that can arise vis-à-vis risks and responsibility as this volume itself will show.

1.3.8 Concluding Responsibility

The above sections illustrated some of the more prominent senses, conceptions, and dimensions of responsibility, but others remain unexplored. The concept of responsibility, just like the concept of risk, remains contested and unexhausted. Responsibility is simply too complex and the literature on the topic is too extensive to allow for exhausting summaries. With this in mind, to varying degrees and, respectively, all the above senses of responsibility will feature in the contributions to this volume. This is why the preceding should serve the reader moving forward.

Discussion so far has focused on the topic of risk and responsibility mostly by attempting to elucidate them separately. In turn, the contributions to this volume bring the two concepts together by exploring their interplay in theories and applications. The following will sketch out the ways in which this exploration will be achieved.

1.4 Risk and Responsibility in Context

The link between risk and responsibility is evident at least in virtue of the fact that many present-day risks are construed as something that should be managed, avoided, mitigated, controlled, or, in some other sense, addressed. Thus, the idea that one (be it an individual or a collective) ought to do something with regard to risks carries with it the notion of responsibility: there are risks for which one should take responsibility, be held responsible for doing or failing to do so, or be held responsible for creating in the first place. This is far from tying risk and responsibility under one conceptual umbrella or endorsing any particular theory of the concepts, respectively. However, this opens the door toward exploring the manifold connections between risk and responsibility as achieved in this work.

The contributions to this volume approach the topic of risk and responsibility from a variety of perspectives. They grapple with, clarify, and expand the relationships between risk and responsibility across various philosophical areas. In so doing, both concepts retain their complexities, nuances, and variations and, at the same time, manage to capture some of the most pressing and difficult moral challenges of our modern world.

In *Part I: Conceptual Context*, the contributions focus primarily but not exclusively on broader theoretical issues associated with risk and responsibility. *Ibo van de Poel and Martin Sand's* chapter, "Responsibility beyond Control," examines the control condition of responsibility, which states that it is unreasonable to hold agents responsible for things that are beyond their control. Against the traditional view that sees control as a precondition of responsibility, the authors propose an alternative view that preserves the strong (conceptual) connection between control and responsibility but allows for a reversal of their relation. In the authors' view, responsibility sometimes precedes control because agents can reasonably take responsibility for things that are not yet under our control. The authors also discuss under which conditions it may be reasonable to take responsibility for certain risks beyond our control, and whether it may sometimes be morally required to do so.

In their chapter "Risk Mismanagement: The Illusion of Control in Indeterminate Systems," *Benjamin Hale and Kenneth Shockley* argue that findings from social choice and game theory, which advance the view that many outcomes are not merely uncertain but indeterminate, complicate the epistemic and metaphysical picture that informs risk-oriented views. Such views imagine the future as unfolding according to a set of risks that are epistemically available through modeling and projection. However, the authors argue that approaches to risk management that ignore indeterminacy result in framing that distorts our decision options, our sense of what is feasible, and the range of our responses. Not only do the authors reveal and explain such distorting effects, but they also advance a framework that better reflects our social realities.

In *Part II: Legal Context*, the relationship between risk and responsibility as it manifests in the context of law is examined. *RA Duff's* chapter, "Risk, Responsibility, and Pre-Trial Detention," discusses the justification of pre-trial detention. Imprisoning people who have not yet been convicted but are awaiting trial is justified by preventive reasoning. Such reasoning regards pre-trial detention as necessary to avert various risks: that the defendant will fail to appear for trial, interfere with witnesses, or commit other kinds of offence if left free. The author argues that this kind of justification seems inconsistent with the presumption of innocence and the liberal principle that the state should respect the freedom and autonomy of responsible citizens. Duff rejects some current attempts to justify pre-trial detention and offers instead a plausible alternative. This is based on the distinctive responsibilities that define the role of the criminal defendant and can support imposing special constraints, even preventive detention, on those awaiting trial.

Anne Ruth Mackor's chapter, "Risks of Incorrect Use of Probabilities in Court and What to Do about Them," investigates the risks involved in the judicial interpretation and application of probability statements. Probabilities play an important role in the proof of facts in trials. However, judges are not trained in probabilistic reasoning, which leads to errors. The author argues that more education in probability theory is not enough and a more radical solution might be needed, such as the introduction of "probability-judges" in evidentially complex cases. Probability judges are experts in probability theory who sit in mixed chambers of the court (i.e., chambers that are composed of judges and probability experts). Such a solution would not conflict with the rule of law, the fundamental right to a fair trial, nor would it open the floodgates to other expert judges and mixed chambers. Nevertheless, the author acknowledges that empirical testing is needed to find out whether such mixed chambers help to reduce the risks of flawed probabilistic reasoning.

Part III: Bioethical Context takes up risk and responsibility in the context of health and bioethics. In "The Failure of Luck Anti-egalitarianism," *Sven Ove Hansson* criticizes the so-called "luck-egalitarian" view that a person deemed responsible for her own disease or injury should be deprived of healthcare resources. According to luck egalitarianism, society should make up for misfortunes that are due to brute luck (i.e., the result of risks that are not deliberate gambles, such as misfortunes in genetic makeup), but it should not compensate for disadvantages that are down to option luck, which are a matter of one's own risk-taking. Hansson shows that conditions that would make the luck-egalitarian claim plausible cannot be fulfilled. The conditions identified and found wanting are the following: (1) that it can be determined whether a person caused her own disease or injury, (2) that blame responsibility can justifiably be assigned to her if she did so, and (3) that this is a sufficient moral reason for withholding treatment that would otherwise have been available to her. In light of such failings, and the fact that the luck-egalitarian position leaves the privileged unaffected while punishing the poor, Hansson argues that luck egalitarianism is a misnomer. This position should be called "luck anti-egalitarianism."

In "Moral Responsibility and Public Health Risks: Examples from the Corona Pandemic," *Jessica Nihlén Fahlquist* addresses the coronavirus pandemic as an example of the relevance of responsibility to public health risks. Fahlquist argues that the pandemic has given rise to a number of ethical questions. Against the backdrop of different conceptions of moral responsibility, the author investigates some of these questions. In particular, the author focuses analysis on how individual responsibility and governmental responsibility ought to be conceived, as well as on how responsibility ought to be distributed in the pandemic.

In "Responsible Risking, Forethought, and the Case of Human Gene Editing," *Madeleine Hayenhjelm* provides an account of responsible risking. After discussing the concepts at stake, the author investigates responsible risking by focusing on various conditions that this notion entails, as well as the ethical debate on human germline gene editing. The author reveals a host of epistemic concerns, as well as a special category of potential losses that are in principle incompensable in the germline gene editing case. The author argues that responsible risking involves at a minimum the avoidance of such risking unless there are extraordinary reasons to do otherwise.

Part IV: Technological Context focuses on risk and responsibility across a range of technological applications and challenges. In "Emotions, Risk and Responsibility: Emotions, Values and Responsible Innovation of Risky Technologies," *Sabine Roeser* and *Steffen Steinert* focus on the contribution that emotions and values can make to the responsible innovation of risky technologies. The authors develop the idea that emotions can play an important role in ethical decision-making about risky technologies by expanding its range of application to the following key stakeholders: universities, industry, policy makers, and the public. They advance a position that supports the view that embedding emotions and values in the innovation of risky technologies can enhance the quality of deliberation and decision-making regarding technological risks, help to overcome stalemates, and lead to technological innovations that are morally and socially acceptable and responsible.

Sven Nyholm's chapter, "Responsibility Gaps, Value Alignment, and Meaningful Human Control over Artificial Intelligence," investigates four different kinds of responsibility gaps. A responsibility gap occurs when some event or outcome is such that it would be fitting to hold somebody responsible for it, but there is no one who could fittingly be held responsible. The author focuses on forward-looking positive responsibility gaps and relates these to the so-called value alignment problem in AI ethics. This is the problem of ensuring that advanced AI aligns with human values, interests, or aims, so that risks related to advanced AI are mitigated. Nyholm criticizes some recent proposed solutions to this problem by, *inter alia*, exposing the difficulties with implementing them into practice and in relation to real-world risks (e.g., risks related to advanced AI).

In "Radioactive Waste and Responsibility toward Future Generations," *Céline Kermisch* and *Christophe Depaus* discuss the responsibility toward future generations in light of risks from radioactive waste, which spread over long periods of time. The authors analyze institutional responses that seek to address this, such as those from the International Atomic Energy Agency (IAEA), the Nuclear Energy Agency (NEA), the International Commission on Radiological Protection (ICRP), and the European Union (EU). Deep geological disposal, which is the technical solution that benefits from international consensus, is considered. The authors criticize the implementation of retrievability in light of our responsibility toward future generations and show that, if unrestrained, retrievable geological disposal is far from the obvious ethical choice.

In *Part V: Environmental Context*, the contributors assess the risk-responsibility relationship vis-à-vis challenges and problems within the context of climate change. *Neelke Doorn* and *Samantha Copeland*, in their chapter "Resilience and Responsibilities: Normative Resilience for Responsibility Arrangements," offer a critical review of various definitions and conceptions of resilience. The authors advance a conceptualization of resilience that, among other things, integrates responsibility. Their explicitly normative notion of resilience can account for the responsibilities of different actors in realizing resilience (i.e., their task responsibilities). The concept of risk is operative within this conception because the resilience at stake for the authors is against changing situations that represent increased risks to the functioning of systems.

In "Individual Climate Risks at the Bounds of Rationality," *Avram Hiller* discusses the moral appropriateness of disregarding small risks. This is a hugely important topic not only because it can be said that all ordinary decisions involve *some* risk but also because many mundane actions performed by individuals (e.g., driving) can be said to contribute to climate harms. Hiller argues that because our rationality is *bounded*, it is not possible for us to include every small risk in our decision-making process and heuristics may be reasonably used. However, contra some thinkers, Hiller argues that this does not violate the spirit of expected value theory but rather shows that we should adopt a so-called *two-level* view. As for individual climate-related risks, Hiller argues that the use of heuristics does *not* permit the *general* ignoring of climate-change-related risks by individuals.

1.5 Conclusion

As the descriptions of this chapter reveal, this volume's contributions engage with risk and responsibility as a theoretical and practical topic, where conceptual issues come to the fore in various contexts. The common link among the chapters is their analysis of risk and responsibility and the interplay between these concepts. However diverse and distinct the volume's contributions, they all highlight challenges, problems, and potential opportunities to address risk and responsibility as a theoretical and practical topic of philosophical analysis. The thematic unity of respective parts of the volume is drawn together as a whole through the conjunction of risk and responsibility. This conjunction underlines the relative novelty of this volume and its constitutive chapters; it explores the multifaceted aspects of risk *and* responsibility rather than aspects of risk *or* responsibility. What is more, it shows how risk and responsibility through their application in different contexts play an important role in the contemporary world.

Notes

[1] For example, a loose definition of risk is unacceptable in science, but perfectly acceptable in a friendly conversation.

[2] Renn (1992) divides risk approaches into seven categories belonging to different fields: (1) the actuarial approach, (2) the toxicological and epidemiological approach, (3) the engineering approach, (4) the economical approach, (5) the

psychological approach, (6) social theories of risk, and (7) cultural theory of risk (Renn 1992, 56).

[3] For a useful introduction, see Ian Hacking, *An Introduction to Probability and Inductive Logic* 23–78 (2001).

[4] For useful overviews, see Jonathan Cohen, *An Introduction to the Philosophy of Induction and Probability* (1989) and Donald Gillies, *Philosophical Theories of Probability* (2000).

[5] See, de Finetti, B. (1980). "Foresight. Its Logical Laws, Its Subjective Sources." In H. E. Kyburg, Jr. and H. E. Smokler (eds.), *Studies in Subjective Probability*. Robert E. Krieger Publishing Company (Original work published 1937).

[6] Classic sources are in Reichenbach (1949) and von Mises (1957).

[7] The standard source is Venn (1876). However, the appellation is found in Reichenbach (1949, 374). See Hájek (2007).

[8] See Aven (2003, xii), who refers to probability and probability calculus as "the sole means for expressing uncertainty."

[9] See Ramsey (1931), de Finetti (1937), von Neumann and Morgenstern (1944), and Savage (1954/1972).

[10] See Hansson (2003) for detailed criticism of various ethical theories' ability to contend with risk through the analysis of the mixture appraisal problem.

[11] The disputes about free will and determinism will not appear in this chapter, and the assumption moving forward is that people can be morally responsible. More specifically, this chapter stipulates that moral responsibility is possible because relevant conditions of free will and control can exist.

References

Asveld, Lotte, and Sabine Roeser (Eds.). 2009. *The Ethics of Technological Risk*. London: Earthscan.

Aven, Terje. 2003. *Foundations of Risk Analysis: A Knowledge and Decision-Oriented Perspective*. Chichester: Wiley.

Baatz, Christian. 2014. "Climate Change and Individual Duties to Reduce GHG Emissions." *Ethics, Policy and Environment* 17 (1): 1–19.

Beck, Ulrich. 1992. *Risk Society: Towards a New Modernity*. London: Sage Publications.

Berkey, Brian. 2014. "Climate Change, Moral Intuitions, and Moral Demandingness." *Philosophy and Public Issues* 4 (2): 157–89.

Bovens, Mark. 1998. *The Quest for Responsibility: Accountability and Citizenship in Complex Organisations*. Cambridge: Cambridge University Press.

Boxer, Karin. 2014. "Hart's Senses of 'Responsibility." In *Hart on Responsibility*, edited by Christopher Pulman, 30–46. London: Palgrave-McMillan.

Bradbury, Judith A. 1989. "The Policy Implications of Differing Concepts of Risk." *Science, Technology, & Human Values* 14 (4): 380–99.

Braham, Matthew, and Martin Van Hees. 2012. "An Anatomy of Moral Responsibility." *Mind* 121 (483): 601–34.

Broome, John. 2012. *Climate Matters: Ethics in a Warming World*. New York: W. W. Norton & Company.

Buchak, Lara. 2014. *Risk and Rationality*. Oxford: Oxford University Press.

Buckareff, Andrei, Carlos Moya, and Sergi Rosell. 2015. "Introduction." In *Agency, Freedom, and Moral Responsibility*, edited by Andrei Buckareff, Carlos Moya and Sergi Rosell, 1–9. New York: Palgrave McMillan.

Cane, Peter. 2002. *Responsibility in Law and Morality*. Oxford: Hart Publishing.

Caney, Simon. 2009. "Climate Change and the Future: Discounting for Time, Wealth, and Risk." *Journal of Social Philosophy* 40 (2): 163–86.

Cohen, Jonathan. 1989. *An Introduction to the Philosophy of Induction and Probability*. Oxford: Clarendon Press.

Copp, David. 2007. "The Collective Moral Autonomy Thesis." *Journal of Social Philosophy* 38 (3): 369–88.

Corlett, Angelo J. 2006. *Responsibility and Punishment*. 3rd ed. Dordrecht: Springer.

Cripps, Elizabeth. 2013. *Climate Change and the Moral Agent: Individual Duties in an Interdependent World*. Oxford: Oxford University Press.

De Finetti, Bruno. 1937. "La prévision: ses lois logiques, ses sources subjectives." *Annales de l'institut Henri Poincaré* 7 (1): 1–68.

Erman, Eva, and Niklas Möller. 2018. "The Interdependence of Risk and Moral Theory." *Ethical Theory Moral Practice* 21: 207–16.

Fischer, John M., and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (23): 829–39.

French, Peter A. 1984. *Collective and Corporate Responsibility*. New York: Columbia University Press.

Fruh, Kyle, and Marcus Hedahl. 2013. "Coping with Climate Change: What Justice Demands of Surfers, Mormons, and the Rest of Us." *Ethics, Policy and Environment* 16: 273–96.

Gilbert, Margaret. 1989. *On Social Facts*. London: Routledge.

Gillies, Donald. 2000. *Philosophical Theories of Probability*. London: Routledge.

Ginet, Carl. 2007. "An Action Can Be Both Uncaused and Up to the Agent." In *Intentionality, Deliberation, and Autonomy: The Action-Theoretic Basis of Practical Philosophy*, edited by Sandro Nannini and Christoph Lumer, 243–55. Farnham: Ashgate.

Goodin, Robert E. 1995. *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Hacking, Ian. 2001. *An Introduction to Probability and Inductive Logic*. Cambridge: Cambridge University Press.

Hájek, Alan. 2007. "The Reference Class Problem Is Your Problem too." *Synthese* 156 (3): 563–85.

Haji, Ishtiyaque. 1998. *Moral Appraisability: Puzzles, Proposals, and Perplexities*. Oxford: Oxford University Press.

Hansson, Sven Ove. 2003. "Ethical Criteria of Risk Acceptance." *Erkenntnis* 59 (3): 291–309.

Hansson, Sven Ove. 2004. "Philosophical Perspectives on Risk." *Techné: Research in Philosophy and Technology 8* (1): 10–35.

Hansson, Sven Ove. 2009. "Risk and Safety in Technology." In *Handbook of the Philosophy of Science, Philosophy of Technology and Engineering Sciences*, edited by Anthonie Meijers, 1069–102. Amsterdam: Elsevier.

Hart, Herbert Lionel Adolphus. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press.

Hayenhjelm, Madeleine, and Jonathan Wolff. 2011. "The Moral Problem of Risk Impositions: A Survey of the Literature." *European Journal of Philosophy* 20 (S1): E26–51.

Hiller, Avram. 2011. "Climate Change and Individual Responsibility." *The Monist* 94 (3): 349–68.

Kermisch, Céline. 2012. "Risk and Responsibility: A Complex and Evolving Relationship." *Science and Engineering Ethics* 18 (1): 91–102.

Kutz, Christopher. 2012. "Responsibility." In *The Oxford Handbook of Jurisprudence and Philosophy of Law*, edited by Jules L. Coleman, Kenneth Einar Himma, and Scott J. Shapiro, 548–587. Oxford: Oxford University Press.

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin Company.

Ladd, John. 1991. "Bhopal: An Essay on Moral Responsibility and Civic Virtue." *Journal of Social Philosophy* 32: 73–91.

Lepora, Chiara, and Robert E. Goodin. 2013. *On Complicity and Compromise*. Oxford: Oxford University Press.

Lewis, H.D. 1948. "Collective Responsibility." *Philosophy* 23 (84): 3–18.

Lichtenstein, Sarah, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. 1978. "Judged Frequency of Lethal Events." *Journal of Experimental Psychology: Human Learning and Memory* 4 (6): 551–78.

McKenna, Michael. 2008. "Putting the Lie on the Control Condition for Moral Responsibility." *Philosophical Studies* 139 (1): 29–37.

Mele, Alfred R. 1995. *Autonomous Agents: From Self-Control to Autonomy*. Oxford: Oxford University Press.

Miller, Dale E. 2010. *J.S. Mill*. Cambridge: Polity.

Möller, Niklas. 2012. "The Concepts of Risk and Safety." In *Handbook of Risk Theory*, edited by Sabine Roeser, Rafaela Hillerbrand, Martin Peterson and Per Sandin, 55–85. Dordrecht: Springer.

Narveson, Jan. 2002. "Collective Responsibility." *Journal of Ethics* 6: 179–98.

Oberdiek, John. 2017. *Imposing Risk: A Normative Framework*. Oxford: Oxford University Press.

Oshana, Marina A. L. 2015. Responsibility: Philosophical Aspects. In *International Encyclopedia of the Social & Behavioral Sciences*. 2nd ed., edited by James D. Wright, 587–91. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.63077-5

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.

Placani, Adriana. 2017. "Ethical Dimensions of Risks: Beck and Beyond." In *Law in the Risk Society*, edited by John Fanning and Bald De Vries, 101–15. The Hague: Eleven International Publishing.

van de Poel, Ibo, and Jessica Nihlén Fahlquist. 2012. "Risk and Responsibility." In *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, edited by Sabine Roeser, Rafaela Hillerbrand, Martin Peterson and Per Sandin, 877–907. Springer.

van de Poel, Ibo, Jessica Nihlén Fahlquist, Neelke Doorn, Sjoerd Zwart, and Lambèr Royakkers. 2012. "The Problem of Many Hands: Climate Change as an Example." *Science and Engineering Ethics* 18 (1): 49–67.

Pritchard, Duncan. 2015. "Risk." *Metaphilosophy* 46 (3): 436–61.

Ramsey, Frank. 1931. "Truth and Probability." In *The Foundations of Mathematics and Other Logical Essays*, edited by Richard Bevan Braithwaite, 156–98. London: Routledge & Kegan Paul.

Raterman, Ty. 2012. "Bearing the Weight of the World: On the Extent of an Individual's Environmental Responsibility." *Environmental Values* 21 (4): 417–36.

Raz, Joseph. 2010. "Responsibility and the Negligence Standard." *Oxford Journal of Legal Studies* 30 (1): 1–18.

Reichenbach, Hans. 1949. *The Theory of Probability*. Berkeley: University of California Press.

Renn, Ortwin. 1992. "Concepts of Risk: A Classification." In *Social Theories of Risk*, edited by Sheldon Krimsky and Dominic Golding, 53–79. New York: Praeger.

Roeser, Sabine. 2006. "The Role of Emotions in Judging the Moral Acceptability of Risks." *Safety Science* 44 (8): 689–700.

Roeser, Sabine. 2007. "Ethical Intuitions about Risks." *Safety Science Monitor* 11 (3): 1–13.

Savage, Leonard J. 1954/1972. *The Foundations of Statistics*. 2nd ed. New York: Dover.

Shrader-Frechette, Kristin. 1991. *Risk and Rationality: Philosophical Foundations for Populist Reforms*. Berkeley: University of California Press.

Sinnott-Armstrong, Walter. 2005. "It's Not My Fault." In *Perspectives on Climate Change: Science, Economics, Politics, Ethics*, edited by Walter Sinnott-Armstrong and Richard Howarth, 285–307. Bingley: Emerald.

Slovic, Paul. 2000. *The Perception of Risk*. London: Earthscan.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. 2002. "Rational actors or rational fools: Implications of the effects heuristic for behavioral economics." *Journal of Socio-Economics* 31 (4): 329–42.

Stevens, S.S. 1951. *Handbook of Experimental Psychology*. New York: John Wiley.

Talbert, Matthew. 2016. *Moral Responsibility*. New York: Polity Press.

Thompson, Paul B., and Wesley Dean. 1996. "Competing Conceptions of Risk." *RISK: Health, Safety, Environment* 7 (4): 361–84.

Thomson, Judith Jarvis. 1986. *Rights, Restitution and Risk: Essays in Moral Philosophy*. Cambridge: Harvard University Press.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31.

Venn, John. 1876. *The Logic of Chance*. New York: Dover Publications.

Von Mises, Richard. 1957. *Probability, Statistics and Truth*. London: Allen and Unwin.

Von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Wieland, Jan Willem. 2017. "Introduction: The Epistemic Condition." In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 1–28. New York: Oxford University Press.

Williams, Garrath. 2008. "Responsibility as a Virtue." *Ethical Theory and Moral Practice* 11: 455–70.

Wolff, Jonathan. 2006. "Risk, Fear, Blame, Shame and the Regulation of Public Safety." *Economics and Philosophy* 22 (3): 409–27.

Zimmerman, Michael J. 1988. *An Essay on Moral Responsibility*. Totowa: Rowman & Littlefield.

Zimmerman, Michael J. 2016. "Moral Responsibility and the Moral Community: Is Moral Responsibility Essentially Interpersonal?" *Journal of Ethics* 20: 247–63.

Part I

Conceptual Context

2 Responsibility beyond Control

Ibo van de Poel and Martin Sand

2.1 Introduction

Our modern highly technological society seems to be confronted by a paradox of control. On the one hand, we can – at least collectively – increasingly control our environment, nature, society, and ourselves through new technological means. On the other hand, this very development seems to have led to an increase in humanly-induced technological and natural risks that we cannot, or hardly, control; think of climate change and COVID-19, or of the potential perils of new technologies like climate engineering and synthetic biology. In particular, in relation to autonomous technologies, Ezio di Nucci has recently argued that these technologies are employed in order to gain more control over traffic safety or military operation. As a result, he argues, we have to cede control: "In order to increase (or improve) control, we must cede it, and this is what I argue is paradoxical. [...] The reason for this is simple enough: software – whether it is installed on a car or, as we will see shortly, on many other things – is better than we are at controlling, so that if we really care about control, we must let software take care of it for us – and not just for software or cars" (Di Nucci 2021: xiv).

These new risks emerging from increased control raise profound questions about responsibility. Who, if anyone, is responsible for them? In some cases, like climate change, it seems obvious that we are at least collectively responsible while it remains unclear how such collective responsibilities translate into individual responsibilities. For sure, individuals have some responsibilities and obligations with respect to climate change, like seeing to it that collective agreements to abate it are reached (van de Poel et al. 2012) or making reasonable individual contributions (Björnsson 2021). However, it is eccentric to assume that everyone is *individually* responsible for the whole of excessive climate change, as that is obviously beyond individual control.

In the case of risks of new technologies, like climate engineering, synthetic biology, and artificial intelligence, we may also lack knowledge

DOI: 10.4324/9781003276029-3

about the exact risks. That is to say, the risks may be uncertain or even unknown; we may not only not know the risks beforehand, but even be unable to come to know them before they have occurred (van de Poel 2017). This lack of control, again, raises the question to what extent we can individually and collectively be responsible for such risks.

The above considerations raise some serious questions about the relation between control and responsibility. It is commonly assumed – and echoed by many philosophers[1] – that we cannot be responsible for things beyond our control. Of course, we can inquire what type of control is exactly required for responsibility, or, more precisely, for what type of responsibility like blameworthiness, accountability, or forward-looking responsibility. However, it seems unfair to attribute responsibility to some agent i for some φ if i had no control over φ.

Still, there also appear to be cases in which people take responsibility for something that is at least initially beyond their control. One may think of Greta Thunberg, Nelson Mandela, or Martin Luther King who committed themselves to a greater cause. Typically, these people take a forward-looking responsibility to correct some evil in the world (like abating hunger or injustices), or to see to it that some future risk or hazard (like a climate disaster) is forestalled. They are typically unconcerned about their range of control when taking responsibility; they somehow feel they should take responsibility. This does not mean that they ought to take such responsibility from a moral point of view. Often, this is not the case, and we are usually inclined to judge their responsibility-taking as morally supererogatory (i.e., as morally praiseworthy but not required).

In line with such cases, we will suggest that at least under some conditions it can be permissible and reasonable to take on new forward-looking responsibilities, even if the object of such responsibilities is initially beyond our control. This is not to say that control is irrelevant in these situations, quite the contrary. We will suggest that in these cases, the typical relation between control and responsibility is reversed. Rather than being responsible for what is already under our control (and, perhaps, because it is under our control), we are sometimes moved by the call of responsibility, and as a result of taking responsibility, we aim to increase control.

Our aim in this contribution is to further tease out this idea. In order to do so, we first inquire what type of control is required for responsibility. Since most of the philosophical literature on control and responsibility has focused on backward-looking responsibility, and more specifically on blameworthiness, we start out with a delineation of the control condition for blameworthiness by building strongly on Fischer's and Ravizza's (1998) theory of moral responsibility. Next, we show how these insights translate to the case of forward-looking responsibility and spell out what control would be required for forward-looking responsibility. We then argue that we need to distinguish between being forward-looking responsible for some φ (or others attributing such responsibility to us) and cases in which an agent *takes* forward-looking responsibility. We argue that the latter type of case allows room for taking forward-looking responsibility for things that are still beyond the agent's control. Next, we discuss whether taking on such responsibilities is merely morally supererogatory or not, and whether we can also assume "too much" responsibility. Lastly, we argue that one might understand the reciprocal relation between control and responsibility by zooming in on the underlying notion of moral agency.

2.2 What Control Is Needed for Responsibility?

Control has mainly been discussed as a condition for backward-looking responsibility and more specifically for blameworthiness in the philosophical literature. We take blameworthiness here to mean that it is appropriate to blame an agent i for some action or state-of-affairs φ. So understood, blameworthiness means that i is a proper target for blame (with regard to φ), but it does not necessarily mean that i is also actually blamed, or that it would necessarily be obligatory, or even desirable, to blame i for φ (Sand and Klenk 2021).[2] We also leave open the possibility that if i is not blameworthy in the responsibility-sense here intended, it might nevertheless be possible to appropriately blame her on other grounds.[3]

Blameworthiness is not the only sense of backward-looking responsibility. We may, for example, also distinguish accountability and liability (e.g., van de Poel, Royakkers, and Zwart 2015). Here, we will focus on blameworthiness as the main type of backward-looking responsibility, which is at the fore of much of the recent philosophical literature. Roughly, the idea is that for an agent i to be blame-responsible for some φ, there needs to be a certain connection between i and φ. The question is what minimal conditions need to be met to make it appropriate (or fair) to blame i for φ.

There are several conditions that the connection (between i and φ) may have to meet (e.g., foreseeability), but a necessary condition in any theory of responsibility seems to be *control*.[4] The basic idea is that without some control over φ by i, it would be inappropriate to blame i for φ. Originating from Thomas Nagel's work on moral luck, this intuitive idea has been put into a standard formulation by Dana Nelkin and is known as the control principle (CP): "We are morally assessable only to the extent that what we are assessed for depends on factors under our control" (Nelkin 2013). In a recent publication, one of us (Sand 2020) argued that this formulation of CP is too broad; there are types of moral assessment where control plays no or only a subordinate role. Hence, in the following, we will endorse the alternative and more specific formulation of CP focusing not on moral assessment but responsibility as blameworthiness: "People are blameworthy only for things within their control" (Sand 2020).[5]

But what type of control is exactly required for blame-responsibility? Fischer and Ravizza (1998) argue that what is required is not regulative but guidance control. Regulative control involves the possibility to act otherwise, or to bring about other consequences. Guidance control does not require that; it only requires that the action (or consequence) is in a more limited sense under the control of the agent. For actions, guidance control requires that the action results from a reason-responsive mechanism that is the agent's own. Fischer and Ravizza (1998) argue that the condition of guidance control better meets our intuitions about blame-responsibility in a number of cases than regulative control.

For consequences (or states-of-affairs), the conditions for guidance control are somewhat more complicated. Fischer and Ravizza (1998) distinguish here between what they call consequence-particulars and consequence-universals. Consequence-particulars refer not just to a state-of-affairs but also to the (specific) way in which it was brought about (e.g., "the mayor was killed by me"). For consequence-particulars, they propose the following condition for guidance control: "An agent S has guidance control over a consequence-particular C just in case S has guidance control over some act A, […] and it is reasonable to expect S to believe that C will (or may) result from A" (Fischer and Ravizza 1998, 121).

For consequence-universals, this condition does not hold as Fischer and Ravizza (1998) point out. The reason is that consequence-universals can also be realized by what they call triggering events, like actions by others, or external events. For example, the consequence-universal "the mayor was killed" might be caused by me killing her but also by somebody else doing so or by a natural event like a lightning stroke. In such cases, they argue, guidance control needs to be split between the internal process leading to the action (bodily movement) and the external process from the agent's action to the outcome. For the former, the guidance control conditions for action apply (reason-responsive own mechanism). For the latter, they argue that the agent's action (bodily movement) "must be sensitive … in roughly the following sense: if the actual type of process were to occur and all triggering events that do not actually occur were *not* to occur, then a different bodily movement would result in a different upshot (i.e., … a different consequence-universal)" (Fischer and Ravizza 1998, 112). This condition implies that the outcome of the external process (i.e., the consequence-universal in which we are interested) needs to be *action-responsive* in the right way to the action of the agent.

Later authors have pointed out that this criterion (which is more formally formulated in Fischer and Ravizza [1998]) does not work in some cases in which the actions of different agents *jointly* determine an outcome without having engaged in a joint action (i.e., the actors act independently and unaware of each other) (e.g., Björnsson 2011; Brown 2011). The following is an example of this (the case is called *The Lake* and introduced in Björnsson [2011]): suppose three individuals pour an amount of a substance into a lake, unaware of each other. Two amounts of the substance are enough to poison the lake. Who is responsible for the consequence-universal "the lake is poisoned"?

If we apply Fischer's and Ravizza's criterion, it would seem none of them. The actual type of (external) process here is that all three pour an amount of substance and, therefore, if one of them would have acted differently the same outcome would still apply (as two amounts are enough to poison the lake). Such responsibility attribution, however, seems wrong. We are inclined to say that all three are equally responsible.[6] To deal with this type of case, we might want to weaken the action-responsive condition.[7] We may, for example, formulate a weaker action-responsiveness condition along the following lines: "There is at least one scenario (possible world) in which whether agent i doing or omitting her action makes a difference for the outcome."[8] Such a scenario factually exists for each of them. Consider, for example, the scenario that agent A and agent B, but not agent C pour their amount, then *in this scenario* the outcome is action-responsive to the actions of both agent A and agent B; and we can formulate similar scenarios for agents A and C, and for B and C.

This new action-responsiveness condition is quite weak, and it might well be possible to imagine other cases that show that it is too weak.[9] The point, however, is that there is a reasonable condition for action-responsiveness between the apparently too strong version of the action-responsiveness condition proposed by Fischer and Ravizza (1998) and this rather weak version of the condition. If this is indeed on the right track, it seems to show something important, namely that, as Fischer and Ravizza (1998) suggest, action-responsiveness is the right kind of control condition for the external process, even if we might not yet be exactly sure how to spell it out.[10]

We conclude then that blame-responsibility minimally requires guidance control and that in the case of consequence-universals this guidance control has two components, namely an internal reason-responsive mechanism that is the agent's own and which results in the action of the agent, and an external process that is action-responsive (i.e., the consequence-universal needs to be action-responsive in the right sense to the agent's action).

2.3 Control and Other Types of Backward-Looking Responsibility

Our aim now is to investigate whether the established control condition also applies to forward-looking responsibility. Before we do so, it is worthwhile to briefly consider the question whether the previously discussed condition of guidance control also applies to other kinds of backward-looking responsibility besides blameworthiness. We will briefly consider two other main types here, namely accountability and liability.

We take it that if we hold an agent i accountable-responsible for some φ, in which φ again is an action or outcome, we ascribe an obligation to i to account for the occurrence of φ (or at least i's role in the occurrence of φ). It seems that for such ascription to be appropriate, it would not be required that (we know for sure that) i had control over φ but only that we have a reasonable suspicion (expectation) that i had control over φ.

Take the following simple case: you are having a conversation with someone else and suddenly you slap that person in the face. It would seem completely appropriate for her to ask: why did you slap me in the face? And by asking this, the person demands you to account for what you did. Now, perhaps, you are able to provide an explanation of your action that shows that it was not under your control. Maybe you have a condition that sometimes, unexpectedly, causes seizures of sorts, like slapping others, that is not under your control (because it is not reason-responsive). While this may be a perfectly acceptable explanation, which also shows that it would be inappropriate to blame you, it does not mean that the initial ascription of accountability was inappropriate. On the contrary, by holding you accountable, your counterpart confirms that you are a moral agent, who under normal conditions is able to control herself and hence is responsible for her actions, albeit not for this specific action (cf. Watson 2004, 8).

Something similar may well apply to liability-responsibility, which we take to be the obligation to rectify some φ (for example, by compensation or repair). Some authors hold that you can only be *morally* liable for some φ if you are also blame-responsible for that φ (e.g., Hart 1968). If that were the case, it would follow that you can only be morally liable for things under your (guidance) control. Others hold that sometimes causal responsibility, rather than blame-responsibility, may be enough to be morally liable (e.g., Honoré 1999). Consider again the case of you slapping someone in the face. This time the other person is seriously hurt and in pain. It would seem appropriate to say that you are morally liable in this situation (assuming you have regained control over your actions) to help that person and to call a doctor, for example. Such cases still require some control (i.e., same basic control over one's actions and control over some action that rectify φ), but they do not require past control over the occurrence of φ.

2.4 Forward-Looking Responsibility

Let us now focus on the control condition for forward-looking responsibility (other authors have used somewhat different terms here like "prospective responsibility" or "active responsibility") (e.g., Bovens 1998; Cane 2002). We take forward-looking responsibility to mean that the agent i has an obligation to see to it that φ (with φ a state-of-affairs) (cf. Goodin 1995).[11] While we can talk about both forward-looking and backward-looking responsibility in the past, present, or future tense, what distinguishes the two is that when we ascribe backward-looking responsibility, we do so *from the viewpoint* that φ has already occurred (even when this φ is in the future); we may, for example, ask whether agent i would be backward-looking responsible, if φ were to happen in the future. But in answering this question, we take an imaginary viewpoint at some future moment in time in which φ has already occurred and can no longer be changed. Conversely, if we ascribe forward-looking responsibility, we do so from the point of view that φ has not yet occurred. Of course, we can ask whether an agent i was forward-looking responsible for some φ that happened (or did not happen) in the past, but we should judge the responsibility ascription from the viewpoint that φ has not yet occurred. These distinctions will turn out to be important when it comes to the question what type of control is needed for forward-looking responsibility. They also underline, as did the previous section on accountability and liability, that we cannot simply assume that the control condition applies equally to different kinds of responsibility.

Therefore, to tease out the control condition for forward-looking responsibility, we will start with a rather general characterization of forward-looking responsibility and the type of control that seems required. We have seen that forward-looking responsibility can be understood as the obligation to see to it that φ, from the viewpoint that φ has not yet occurred. In terms of control, this seems to require that the responsible agent has some forward-looking control, or what we may call *causal efficacy*, with respect to φ.

One way in which we may understand such causal efficacy is as the capacity to *ensure* φ. This is, however, a quite strong condition because in order for i to have the capacity to ensure φ, i must be able to realize φ *under all possible external conditions*. Effectively, this means that i should have regulative control over φ.[12] But perhaps there is another plausible way for understanding causal efficacy that does not require regulative control but only guidance control. To see whether that is indeed possible, let's look at what is typically expected from an agent who has forward-looking responsibility for φ.

Goodin (1995, 83) suggests that forward-looking responsibility "require[s] certain activities of a self-supervisory nature from [agent] A. The standard form of responsibility is that A sees to it that X. It is not enough that X occurs. A must also have 'seen to it' that X occurs. 'Seeing to it that X' requires, minimally: that A satisfy himself that there is some process (mechanism or activity) at work whereby X will be brought about; that A check from time to time to make sure that that process is still at work, and is performing as expected; and that A take steps as necessary to alter or replace processes that no longer seem likely to bring about X."

A few things are important here. First, the most important criterion in fulfilling one's forward-looking responsibility is *not* that φ (or X in Goodin's terminology) occurs, but rather that agent i (A in Goodin's terminology) *has seen to* it that φ occurs. Second, it is not required that i brings about φ by an action of her own, it is enough that there is a process P that results in φ and that i has certain abilities with respect to that process P (monitoring it, intervening in it or switching to a process P*). This gives i some discretionary room in deciding how φ is to attain.[13] It is for this reason that it seems proper to conceive of the obligation to see to it as a responsibility rather than as a duty, as duties typically refer to (specific) actions that an agent should do or refrain from (van de Poel 2011).

One consequence of the above is that it seems possible that i has fulfilled her obligation to see to it that φ without φ actually attaining. This also seems in line with intuitions about when it is appropriate to attribute forward-looking responsibility. Consider the following example: it seems appropriate to ascribe the responsibility to see to it that passengers are *safely* transported from A to B to a public transport company, or its director(s). Now, this responsibility, among others, implies that we expect the company director to see to it that qualified drivers are hired, that they are instructed to drive safely, that the company buys safe vehicles, that these vehicles are inspected and maintained regularly, and so forth. In other words, we expect the director to see to it that certain processes are in place that, at least in normal circumstances, would guarantee the safety of the passengers. We typically do not expect, however, the company director to be able to prevent all possible accidents, as there can still be cases like, for example, a storm or a terrorist attack that the company director cannot prevent. We accept, thus, that there are scenarios in which the passengers turn out not to be safe, despite the fact that the director has fully discharged her forward-looking responsibility. And the fact that these cases are beyond the company director's control does not invalidate the ascription of forward-looking responsibility beforehand; it is still perfectly appropriate to say that the company director has a forward-looking responsibility for the safety of the passengers of the company when traveling in the companies' vehicles.

This suggests that in order to appropriately attribute forward-looking responsibility for φ to an agent i, i need *not* be able to ensure φ *in all circumstances*. Rather, we would require that i must be able to ensure φ *under normal circumstances* (van de Poel, Royakkers, and Zwart 2015). We propose the following set of conditions as a first approximation to express this:

1. Agent i knows at least one feasible process P that results in φ
2. i can undertake a set of supervisory activities that allow i to monitor P and to intervene in P (if necessary) so that i can ensure that P occurs and results in φ under normal circumstances

It should be noted that these conditions do not require that there is an alternative process P* that achieves φ and to which agent i can switch if P gets blocked. This is not required as a minimal condition for appropriately attributing forward-looking responsibility. This possibility has emerged since we no longer require that i can ensure φ *in all circumstances*.

The proposed set of conditions, thus, does not require i to have regulative control over φ, but only some form of guidance control. Similarly, to the case of blame-responsibility for consequence-universals, this guidance control has an internal and external component. The internal component is that i should have guidance control over the mentioned set of supervisory actions; this means that these supervisory actions should result from a reason-responsive process that is the agent's own. The external component is that the occurrence of φ should be action-responsive to the exercising of these supervisory actions. In this case, this action-responsiveness is cashed out in terms of the outcome φ being responsive to the monitoring of, and potential intervention in a process P by i. Although this is a somewhat different condition for action-responsiveness than in the case of blame-responsibility, it still is an action-responsiveness condition. Whereas in the case of blame-responsibility, action-responsiveness would minimally require that there is a set of (perhaps counterfactual) circumstances (i.e., in a possible world) in which i can prevent the consequence-universal φ from occurring, in the case of forward-looking responsibility it requires minimally that there is a set of (perhaps counterfactual) circumstances in which i can make φ occur.[14]

2.5 Taking Responsibility

Now that the control condition for forward-looking responsibility has been clarified, we will look at cases in which agents actively take or assume responsibility, rather than being held or ascribed a responsibility by others.

We take the following to be the basic form of any responsibility ascription:

Agent j attributes to agent i the responsibility for φ

Using this scheme, we can understand taking responsibility as a special case of responsibility ascription, namely as the case in which j = i.

However, the conditions under which agents can meaningfully take responsibility for φ are somewhat different, and less strict it would seem, than the conditions under which responsibility can be attributed by other agents.

Another way of phrasing this might be to say that an agent i can attribute responsibility to herself from two different perspectives. The first perspective is the third-person perspective in which i asks what responsibility can reasonably be attributed to her (by others, but perhaps also from a general, moral point of view), and a first-person perspective, from which she asks the question: "what do I feel responsible for?" or "what do I aspire or want to take responsibility for?" While the third-person perspective may set limits on what she should take responsibility for, the first-person perspective creates room for taking more responsibility than what one is strictly required to do.[15]

This seems particularly the case for forward-looking responsibility and control, on which we will focus here. We suggest that we can reasonably take forward-looking responsibility for things not yet under our control, but over which we can reasonably expect to gain (some)[16] control, if we seriously try. This possibility can be illuminated by briefly comparing backward-looking responsibility (and in particular blameworthiness) and forward-looking responsibility again and emphasize the differences to which we alluded earlier. The difference is this: when we ascribe backward-looking blame-responsibility (either to ourselves or to others), we do it from the viewpoint that φ has already occurred. In other words, we do it from the viewpoint that we can no longer execute control over φ (as we cannot change the past). However, this is different in the case of forward-looking responsibility, which we ascribe from the viewpoint that φ has not yet occurred. Things that haven't occurred yet do not automatically fall within the range of anyone's control (e.g., volcanic eruptions). While it may be inappropriate for others to ascribe forward-looking responsibility for things currently beyond our control, it seems that we can reasonably assume such responsibility provided that it is reasonable to assume that we can acquire the required control at some not-too-distant point in the future.[17] This ascription merely presumes that one is in control of being able to obtain the required control over φ in a not too-distant future.

The following example illustrates this idea: suppose someone is worried about traffic safety in her neighborhood. She is aware of a number of possible measures that can improve the situation, like the placement of traffic lights, the lowering of speed limits, speed bumps, or other road reconstruction measures. However, she lacks the control over the introduction of such measures that are required to see to it that the traffic situation is reasonably safe in the neighborhood. In this situation, it would clearly seem unreasonable to ascribe a forward-looking responsibility to her to see to it that the traffic situation is reasonably safe in her neighborhood. This responsibility, so it seems, should be attributed to the relevant civil servants or perhaps to the city council. Nevertheless, it is conceivable that they fail to act and that she is so (morally) upset by the situation that she decides to take responsibility for seeing to it that the traffic situation becomes reasonably safe. Despite initially lacking the control to exercise that responsibility, she may look for ways to acquire such control, e.g., placing warning signals, organizing the neighborhood, or running to be elected into the city council.

As this example suggests, taking responsibility may be considered rational and reasonable under a set of conditions like the following:

- i reasonably believes that she has, or can acquire knowledge of, at least one feasible process P (mechanism, causal pathway) resulting in φ
- There is a set of supervisory actions A through which i can monitor and intervene in P so that
  - The occurrence of φ (through P) is action-responsive to A
  - i reasonably believes that she has or can acquire guidance control over A

While the conditions are somewhat similar to the case in which forward-looking responsibility is ascribed from a third-person perspective, there are two important differences. The first and most important difference is that in taking responsibility the agent does not already need to have the required control but only needs to reasonably believe that she can acquire the required control. Secondly, and related to this, it seems that in the case of ascribing responsibility to others, we typically attune the responsibility that we can reasonably ascribe to an agent i to the control i already has, while in the case of taking responsibility, the responsibility seems to come first, and we then attune the required control in order to be able to fulfill that responsibility.

2.6 When Should People Take Forward-Looking Responsibility for Things beyond Their Control?

Taking forward-looking responsibility for something that is beyond one's control may be seen as a voluntary commitment. This suggests that taking such responsibilities is, at least usually, not morally required. Moreover, in many cases, it would seem morally praiseworthy to take on new responsibilities. This suggests that assuming such responsibilities is morally supererogatory (van de Poel and Sand 2021). This, however, needs to be qualified: situations are conceivable in which it is morally undesirable to take on new responsibilities, as well as situations in which it may be morally required to take on new responsibilities.

In so far as taking responsibility equals a voluntary commitment, its moral status is somewhat similar to that of promising. Promising is in itself not morally good or bad; it very much depends on what is promised. For example, one must not promise to do morally bad things (e.g., to kill somebody for money), nor should one make promises that cannot be kept. But even if certain promises are neither immoral nor unfeasible, there may be reasons why it is (morally) undesirable to make them.

One concern is that promises introduce new obligations, the fulfillment of which may conflict with the fulfillment of other (moral) obligations the agent already has. So even if the new obligations can be fulfilled, the fact that their fulfillment comes at the expense of fulfilling other moral obligations may, at least in some cases, be a reason why one should not make the promises in the first place.

Something similar applies to taking forward-looking responsibility for things beyond our control. Assuming such responsibilities introduces a range of new (moral) obligations for the agents, not just the obligation to see to it that φ, but also an obligation to increase one's span of control so that one can see to it that φ. Acquiring such control may, depending on the case, require quite some efforts on behalf of the agent and therefore conflict with other obligations.

Moreover, increased control – as a result of taking responsibility – may itself introduce new responsibilities, even beyond the responsibilities that were initially taken by the agent. Take the earlier example of the local resident who takes responsibility for traffic safety in her neighborhood. Assume she decides to try to get elected in the city council, and she succeeds; this obviously leads to many new responsibilities beyond the responsibility for traffic safety in her neighborhood, for which she took responsibility.

The more general point is that responsibility and control may mutually reinforce each other. An example of global scale is the attempt to develop geoengineering as a way to mitigate climate change. While such attempts have been criticized in the philosophical literature as a technological fix that undermines the motivation to solve the "real" problem (i.e., too high emission levels) (cf. Gardiner 2010), it may also be interpreted in a more positive light as an attempt to increase humanity's control so that we can collectively better take forward-looking responsibility for mitigating climate change. The worry that this reply to climate change nevertheless raises is that by trying to increase control over the climate, we may well introduce *new uncontrollable* risks. Although it is conceivable that these can eventually also be brought under human control, one might not only worry that this is an endless process but also that somewhere along the road, new risks are introduced that are (clearly) unacceptable.

Another worry that may be raised by the potentially mutually reinforcing dynamics of responsibility and control is that there are things one should accept to be beyond control. We are not thinking here about collective or global issues such as climate change, poverty, and environmental degradation. Rather, on the individual level, there are some things which one should not aspire to control (at least directly) and, hence, should not take responsibility for. There are, for example, limits to the extent to which one not only can, but also should, take responsibility for one's own happiness.[18]

The above should not be interpreted as a plea against taking responsibility. As the examples in the introduction show, there are many situations in which taking responsibility is morally praiseworthy. Nevertheless, there are also situations in which it is not praiseworthy and perhaps even morally undesirable to take on certain new responsibilities.

On the other side of the spectrum, one may wonder whether there are situations in which it is morally obligatory to take on new responsibilities. We suggested earlier that one should at least assume responsibilities that others can reasonably attribute to us. So, even if others do not actually, overtly attribute such responsibilities, we should probably assume them ourselves. Moral responsibility does not need a spokesperson.

However, we have also suggested that such attributable responsibilities are typically limited to what is currently within our control. Still, one might wonder whether others can also not sometimes reasonably or appropriately attribute responsibilities to us for things beyond our control. Alfano and Robichaud (2018) briefly mention an example in which someone (a diplomat or politician) is tasked with the responsibility to solve the Middle-East conflict. Such a political position requires the agent to acquire control in a sense that she usually doesn't have at the moment when she is accepting the task. As suggested by the example, it seems true that others can attribute (or delegate) forward-looking responsibilities to us for things that are beyond our control, but it would also seem that such attributions are only *appropriate* attributions of *moral* forward-looking responsibility, if they are *voluntarily accepted* by the responsible agent. That is to say, the attribution may sometimes be inappropriate because the agent to whom responsibility is attributed (or delegated) lacks the capability to exercise the responsibility (as is obviously the case with regard to Jared Kushner as Alfano and Robichaud [2018] correctly point out). But even when the attribution is not inappropriate, it would only seem to be an attribution of (political, legal) *task responsibility*, not of moral responsibility. It only becomes a moral forward-looking responsibility, if and once the agent to whom this responsibility is attributed voluntarily accepts the responsibility or at least accepts the task that accompanies the responsibility. This type of cases differs from the prototypical case of an agent voluntarily taking forward-looking responsibility for things beyond her control that we discussed before; it is in any cases crucial that the agent *voluntarily accepts* the forward-looking responsibility attributed to her by others for things beyond her control.

Still, there may also be situations in which it is not just praiseworthy but even morally obligatory to take on new responsibilities. Three types of considerations seem to be relevant here (cf. Miller 2001). First, the seriousness and urgency of a certain moral situation. The more serious or urgent the situation, the greater the moral demand for someone to take responsibility for it. Second, the degree to which an agent has or can acquire unique capabilities to address the problem.[19] A third consideration seems to be the agent's current connection with the problem ("connection" is here understood broadly). There may, for example, be cases in which one is (partially) morally blameworthy or morally liable for the problem, which may introduce an obligation to take responsibility for it. While a causal connection alone (without blame or liability) is probably not enough to introduce an obligation to take a responsibility, it may be a factor among the other mentioned considerations. Yet, another way that one may be connected to the problem is that it is in one's realm of authority (e.g., as politician) without necessarily already possessing the required control to solve it.

2.7 Moral Agency

We have suggested that responsibility and control have a reciprocal relation. While in many cases control precedes responsibility and it may be unfair or inappropriate to hold someone responsible for actions or consequences beyond that person's control, in other cases taking (forward-looking) responsibility may precede control and may motivate expanding one's scope of control. Still, there seems to be an important way in which these two types of situations are similar despite their apparent difference. We suggest that in both cases, the relation between responsibility and control suggests a particular notion of moral agency.

As Fischer and Ravizza (1998) point out, attributions of moral responsibility to an agent are historically preceded by that agent having taken responsibility for her actions in a more general sense. With taking responsibility, they do not mean that an agent takes a specific responsibility, as we have used the phrase above. Instead, they mean that humans at some point in their upbringing begin to see their actions as their *own*. At some point in their upbringing, humans take or accept moral authorship for their actions, and the consequences of these actions. This acceptance of moral authorship by an agent is in their view a (historical) precondition for guidance control.[20]

By accepting moral authorship for one's action and their consequences, one typically also starts to conceive of oneself as a proper target of praise and blame, or sanction and reward. In other words, one starts to think of oneself as a being that can properly be held responsible by others, or oneself. A third aspect of moral agency (in addition to accepting moral authorship for one's actions, and conceiving of oneself as a proper target of reactive attitudes) is to start seeing oneself, and being recognized by others, as part of a larger moral community, a community that to some extent shares certain moral norms and values, where it is considered appropriate to hold another accountable for living by these moral norms and values (cf. Kutz 2000).

While becoming a full-blown moral agent may, as Fischer and Ravizza (1998) suggest, historically precede the attribution of specific moral responsibilities, we would like to suggest that the scope of our moral agency, and hence the scope of our moral responsibility, is not given but may change over time. And it may do so in two ways, namely (1) by extending (or reducing) our span of control in the world, we increase (or decrease) the scope of our moral agency in the world and hence the scope of our moral responsibilities, and (2) by (voluntarily) taking on new (forward-looking) responsibilities, we extend our moral agency, and to effectuate that extended moral agency, we may need to increase our scope of control.

From our point of view, the traditional discussion about responsibility has focused only on the first route. It was assumed that control is a precondition for responsibility and that the only way in which our moral agency and responsibility can increase is through a preceding increase in control. However, there is also a second possibility, where we start with (voluntarily) extending our moral agency and hence our responsibility, and as a result of such (voluntary) commitment need to try to extend our scope of control. The existence of such a route is indeed suggested by the fact – laid bare by Fischer and Ravizza (1998) – that all responsibility attributions are grounded in an agent having taken responsibility in a more fundamental and basic sense.

2.8 Conclusion

Traditionally, control is seen as a precondition for responsibility. We have sketched an alternative view. On this view, there is still a strong (conceptual) connection between control and responsibility, but control does not always precede responsibility. Rather, the relation may be reversed. Responsibility might sometimes precede control. The main reason is that we can reasonably take responsibility also for things that are not yet under our control.

Taking responsibility is not only important as a way to acquire specific forward-looking responsibilities, including for things not yet under our control. It is also a more fundamental phenomenon that precedes any appropriate responsibility attribution in a more fundamental sense as Fischer and Ravizza (1998) already suggested. In order for certain actions to be the agent's own and to be under her control, she first needs to accept moral authorship or agency over her actions.

On the picture that arises, moral agency is not something given but something that has been acquired and assumed (typically during upbringing). Moreover, moral agency comes in degrees, and human agents can assume less or more moral agency, with more moral agency not necessarily being better because – as we have seen – taking on new responsibilities is not always desirable or morally permissible.

The sketched view has a number of implications regarding responsibility for the risks of new technologies. It suggests that we can sometimes take responsibility for technological (or other) risks that are still beyond our control. At the same time, it suggests that taking such responsibilities will typically also require the agent to increase her span of control, and that may not necessarily always be good or desirable. Hence, there is a limit to the extent that agents not only can but also should take on new responsibilities.

Acknowledgments

This publication is part of the project ValueChange that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 788321. This publication also contributes to the research programme Ethics of Socially Disruptive Technologies, which is funded through the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

Notes

[1] Thomas Nagel considers the control principle not as a philosophical artefact, but as being deeply rooted in common sense morality: "Prior to reflection it is intuitively plausible that people cannot be morally assessed for what is not their fault, or for what is due to factors beyond their control" (Nagel 1979: 25).

2 There might be all kinds of reasons, including pragmatic ones, why it may not be obligatory or desirable to blame an agent that is blameworthy. We take the attribution of blame-responsibility, thus, to be an attribution of blameworthiness, not an attribution of blame.

3 For example, some form of legal blame (and penalty) may be appropriate also in cases an agent is not morally blameworthy in a responsibility-sense.

4 Not all authors consider control explicitly as a responsibility-condition, but as far as we see most, if not all, assume it implicitly in one way or the other. As Sand (2020) points out, those who reject the control principle (e.g., Hanna 2014) have to develop a theory of blameworthiness that explains why blaming people for random harms or the wrongs of other people is unacceptable, something to which CP has a clear answer.

5 Remarkably few defenses of CP have been developed in the philosophical literature. One of the authors of the present paper defended CP in another publication with an appeal to simplicity (Sand 2020).

6 One might debate whether they are all three (equally) responsible for the "consequence-universal" (that the lake is poisoned) or perhaps for something else (like contributing to the poisoning). It is, in any case, clearly wrong to say that they are not responsible.

7 Björnsson (2011) himself proposes another solution that focuses on whether the actions might *explain* the outcome.

8 This is our proposal, not that of Björnsson (2011) or Fischer and Ravizza (1998), although it is intended to be in line with Fischer's and Ravizza's proposal.

9 A main worry about the weak criterion seems to be that the actual process by which φ is achieved is irrelevant to it, while it intuitively would seem to matter what the actual process was that led to φ. Another worry might be that the criterion cannot distinguish between (relatively) more substantial contributions (like in *The Lake*) and small contributions. Consider, for example, the case of climate change. Here, also a certain threshold of individual contributions needs to be passed in order for the collective (undesirable) effect to occur (although this partly depends on how one exactly understands the relevant physical mechanisms). But, contrary to *The Lake*, much more than two individual contributions are required for the collective effect to occur. If we apply the weak action-responsiveness criterion, climate change – maybe somewhat surprisingly – seems to be under individual control, as for each individual there is at least one scenario in which the contribution of that individual is decisive for whether the threshold is passed or not (depending on how exactly the physical mechanism at play are understood). While in the case of climate change, there may be some individual blameworthiness, it would seem excessive to say that each individual is blameworthy for the total effect (as we are inclined to do in the case of *The Lake*). Perhaps, this needs to be explained by the fact that the cases are different in terms of other responsibility conditions, like wrong-doing.

10 This is not meant to suggest that action-responsiveness exhausts the control condition. Perhaps more is required for control than action-responsiveness (and reason-responsiveness as earlier discussed), like knowledge of the consequences or at least the ability to know the consequences, or – alternatively – one might understand 'knowledge' as an additional condition for proper attribution of moral responsibility, in addition to control.

11 It is not meaningful to talk about forward-looking responsibility for actions, at least for the agent's own actions. Those are better called duties or obligations.

[12] This can be seen as follows. Suppose that there is some process P that leads to φ under normal circumstances. Now also suppose that i has guidance control over P. Now, by having guidance control over P, i also has guidance control over φ (because P would under normal circumstances result in φ). However, such guidance control is not enough to have the capacity to ensure φ because due to external events or actions (i.e., what Fischer and Ravizza call triggering events) something may happen that blocks P or the path from P to φ. Now in order to ensure φ, i should be able to switch to another process P* that also results in φ. This means that agent i should have guidance control over both process P and P* (and perhaps in real-world scenarios over even more processes). Such dual guidance control is effectively a form of regulative control as Fischer and Ravizza (1998) point out.

[13] Interestingly, φ doesn't even have to be brought about (by anyone). It could be a state that is the result of a natural process and forward-looking responsibility ought to ensure that no one is interfering with it.

[14] With 'minimally' we do not mean that these are the conditions under which we can appropriately attribute backward-looking or forward-looking responsibility, but rather that any further specification of the action-responsiveness condition (for appropriate responsibility contributions) should at least be as strong as this minimal criterion. There might in fact be other reasons why ascriptions of forward-looking responsibility are inappropriate. For example, Alfano and Robichaud (2018) suggest that ascriptions of forward-looking responsibility are inappropriate if the standing of the attributer doesn't permit the attribution (e.g., due to lack of authority) or if it overburdens the agent. Overburdening might mean that fulfilling the forward-looking responsibility requires too big of a sacrifice (cf. Fischer and Tognazzini 2011).

[15] Since both perspectives are eventually hers, there can be a misalignment between what she believes her moral obligations to be and what her moral obligations really are. At the same time, she might mistakenly judge her aspirations to be beyond what she is morally obliged to do, while both coincide. In some sense, she can then count herself lucky for doing the right thing (though most likely for the wrong reasons).

[16] How much control is required strongly depends on the context of action and the exact forward-looking responsibility assumed.

[17] A related, yet distinct, idea is that it is sometimes desirable that we take – or at least try to take – responsibility for things that might remain beyond our control. A weaker version of this view is clearly defensible. Whether we get climate change "under control" is currently not predictable, but our chances certainly increase if people give a wholehearted try (an effort that oftentimes motivates others to join). The stronger version is less defensible: in medical situations, when there is literally no way of saving a patient, it is unreasonable to continue with the effort. We thank Adriana Placani for making us aware of this and Sven Ove Hansson for suggesting the formulation "responsibility to try".

[18] As Aristotle already suggested happiness may well be a by-product (i.e., something that is attained in aiming for other things rather than something that can be aimed at or deliberately achieved).

[19] If a choice between agents can be made, it is most reasonable to choose someone who has the relevant control to handle the situation (a doctor to help an accident survivor rather than asking someone, figuring out how to handle a patient).

[20] Remember that guidance control requires that actions originate from a reason-responsive mechanism that is (recognized as) the agent's *own*.

References

Alfano, Mark, and Philip Robichaud. 2018. "Nudges and Other Moral Technologies in the Context of Power: Assigning and Accepting Responsibility." In *The Palgrave Handbook of Philosophy and Public Policy*, edited by David Boonin, 235–48. London: Palgrave MacMillan.

Björnsson, Gunnar. 2011. "Joint Responsibility Without Individual Control: Applying the Explanation Hypothesis." In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Jeroen van den Hoven, Ibo van de Poel and Nicole Vincent, 181–200. Dordrecht: Springer.

Björnsson, Gunnar. 2021. "On Individual and Shared Obligations: In Defense of the Activist's Perspective." In *Philosophy and Climate Change*, edited by Mark Budolfson, Tristram McPherson and David Plunkett, 252–80. Oxford: Oxford University Press.

Bovens, Mark. 1998. *The Quest for Responsibility. Accountability and Citizenship in Complex Organisations*. Cambridge: Cambridge University Press.

Brown, Alexander. 2011. "Moral Responsibility and Jointly Determined Consequences." In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Nicole A. Vincent, Ibo van de Poel and Jeroen van den Hoven, 161–79. Dordrecht: Springer Netherlands.

Cane, Peter. 2002. *Responsibility in Law and Morality*. Oxford: Hart Publishing.

Di Nucci, Ezio. 2021. *The Control Paradox: From AI to Populism*. Lanham: Rowman & Littlefield.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*, *Cambridge Studies in Philosophy and Law*. Cambridge: Cambridge University Press.

Fischer, John Martin, and Neal A. Tognazzini. 2011. "The Physiognomy of Responsibility." *Philosophy and Phenomenological Research* 82 (2): 381–417.

Gardiner, Stephen M. 2010. "Is "Arming the Future" With Geoengineering Really the Lesser Evil?: Some Doubts About the Ethics of Intentionally Manipulating the Climate System." In *Climate Ethics: Essential Readings*, edited by Stephen M. Gardiner, Simon Caney, Dale Jamieson and Henry Shue, 284–312. Oxford: Oxford University Press.

Goodin, Robert E. 1995. *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Hanna, Nathan. 2014. "Moral Luck Defended." *Noûs* 48 (4):683–98. https://doi.org/10.1111/j.1468-0068.2012.00869.x

Hart, Herbert L. A. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Clarendon Press.

Honoré, Tony. 1999. *Responsibility and Fault*. Oxford: Hart.

Kutz, Christopher. 2000. *Complicity: Ethics and Law for a Collective Age*, *Cambridge Studies in Philosophy and Law*. Cambridge: Cambridge University Press.

Miller, David. 2001. "Distributing Responsibilities." *The Journal of Political Philosophy* 9 (4): 453–71.

Nagel, Thomas. 1979. *Mortal Questions*. Cambridge: Cambridge University Press.

Nelkin, Dana K. 2013. "Moral Luck." In *The Stanford Encyclopedia of Philosophy (Winter 2013 ed.)*, edited by Edward N. Zalta. Stanford: Stanford University, Metaphysics Research Lab. https://plato.stanford.edu/entries/moral-luck/

van de Poel, Ibo. 2011. "The Relation between Forward-Looking and Backward-Looking Responsibility." In *Moral Responsibility. Beyond Free Will and Determinism*, edited by Nicole Vincent, Ibo van de Poel and Jeroen Van den Hoven, 37–52. Dordrecht: Springer.

van de Poel, Ibo. 2017. "Society as a Laboratory to Experiment with New Technologies." In *Embedding New Technologies into Society: A Regulatory, Ethical and Societal Perspective*, edited by Diana M. Bowman, Elen Stokes and Arie Rip, 61–87. Singapore: Pan Stanford Publishing.

van de Poel, Ibo, Jessica Nihlén Fahlquist, Neelke Doorn, Sjoerd Zwart, and Lambèr Royakkers. 2012. "The Problem of Many Hands: Climate Change as an Example." *Science and Engineering Ethics* 18 (1):49–68. https://doi.org/10.1007/s11948-011-9276-0

van de Poel, Ibo, Lamber Royakkers, and Sjoerd D. Zwart. 2015. *Moral Responsibility and the Problem of Many Hands*. London: Routledge.

van de Poel, Ibo, and Martin Sand. 2021. "Varieties of Responsibility: Two Problems of Responsible Innovation." *Synthese* 198: 4769–87. https://doi.org/10.1007/s11229-018-01951-7

Sand, Martin, and Michael Klenk. 2021. "Moral Luck and Unfair Blame." *The Journal of Value Inquiry*. https://doi.org/10.1007/s10790-021-09856-4

Sand, Martin. 2020. "A Defence of the Control Principle." *Philosophia*. https://doi.org/10.1007/s11406-020-00242-1

Watson, Gary. 2004. *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.

3 Risk Mismanagement

The Illusion of Control in Indeterminate Systems

Benjamin Hale and Kenneth Shockley

3.1 Introduction

Living in the modern age means standing in relation to the future in a particular way, in a way that sees the future as unfolding according to a set of risks: as hazards and probabilities that are epistemically available for assessment through modeling and projection. But this risk-oriented way of envisioning the future is at once enabling and limiting. It is enabling inasmuch as it empowers us to anticipate the future and prevent bad outcomes; but it is limiting in that it depends on an artificial notion of control and thereby undermines practical deliberation. Indeed, the so-called Risk Society that so enthralled Anthony Giddens and Ulrich Beck in the 1980s and 1990s looms as large today as it did 30 years ago (Beck 1995; Giddens 1984).

Recent observations from social choice and game theory – namely, that many outcomes are not merely uncertain but instead *indeterminate* – complicate the epistemic and metaphysical picture that informs these risk-oriented views (Hale 2019, 2020, 2022; Hardin 2003; Varoufakis 2013). The future itself is heavily dependent upon actions taken by actors in the present. Though policy-makers and ethicists often frame possible futures in terms of decision trees with outputs and probabilities, this is a poor framing. It offers the illusion of control in the face of indeterminate systems and treats the world as if devoid of agency and cooperation. It confuses the model with the reality modeled. Moreover, the usual framing of possible futures resulting from individual versus collective choices misconstrues the social realities that condition not only our deliberation but also the range of future possibilities for addressing the challenges of the modern age.

In this chapter, we consider the ways in which risk orientation colonizes our decision options, distorting our sense of what is feasible and restricting the range of possible responses to complex societal challenges. Namely, many theorists working on contemporary social problems instinctively reduce questions of power and responsibility to the individual or the

DOI: 10.4324/9781003276029-4

collective, effectively neglecting centuries of efforts at coordinating action and militating against more coordinated responses. The false choice between individual and state responsibility obfuscates the possibility of more realistic and effective institutional and non-state cooperative arrangements.

The thesis we develop responds to the tendency to view questions of justice and responsibility in either exclusively individualized or state-centric terms. The increasing appeal of non-state actors in a range of arenas – from COVID to climate change – points to the importance of rethinking the false individual-collective dichotomy. Forcing deliberation into an individual-or-collective framing constrains the available options by failing to acknowledge the indeterminate nature of intervening agential involvement. Ignoring the complex realities of our social environment blinds us to the un-anticipatable futures that these models claim to forecast. In this work, we propose tackling the distorting effects of this approach to risk and instead introduce a framework that both better reflects our social realities and considers a wider range of reasons that inform practical decisions.

This problem applies, of course, to many more issues than the pandemic. From climate change to terrorism, from gun control to abortion, from animal rights to immigration, the problem of risk and indeterminacy looms large. Climate change is the area in which we (the authors) have been writing for most of our careers, but we feel that stepping back from the wicked nature of the climate problem to assess indeterminacy in another context, in this case the COVID context, can be instructive. The sociology, political science, economics, epidemiology, climate science, and management literature have been grappling with risk and uncertainty for 30+ years, mostly in a positivist or descriptivist vein, essentially looking either to shore up the methods of risk assessment or to make sense of the ways in which a risk orientation has shaped modern civilization. Ethics and political theory, considerably more normative branches of inquiry, have often borrowed uncritically from this descriptivist literature to assign responsibility for outcomes and guide action. In this way, the normative community has largely failed to recognize the extent to which the ethical discussion has been framed and, in many, respects the conclusions prewritten, through this operationalization/systemization process of the descriptive sciences. Fortunately, recent work has slowly begun to peel back the layers of indeterminacy in the risk modeling literature and decouple responsibility from risk assessment.

3.2 COVID and the Risk Society[1]

The past two years of COVID-19 have led individuals and states to make a range of decisions based on the assessment of risks. These risks involve concerns about how individual and collective behavior will shape the COVID-modified world in which we live. We have seen risk language enter the discussion throughout the pandemic. First, we were regaled with models of disease spread through populations, then projections of how many would die, then questions about how to "flatten the curve." Later we saw models related to the efficacy of lockdown policies, of mask mandates, of the vaccination rollout. We've seen models pitting losses in the economic sector against losses in health. We've seen philosophers and economists invoke QALYs (Quality Adjusted Life Years) and DALYs (Disability Adjusted Life Years) to argue for and against lockdowns, opening bars, closing schools, imposing curfews, vaccinating the younger population, and so on (Bramble 2020; Jaziri and Alnahdi 2020; Reddy 2020; Rodger and Blackshaw 2022). We've seen armchair epidemiologists roll up their sleeves, conduct their own research, and push for opening schools based on apparent risk to children, ignoring all the while massive methodological and socio-political complications assumed in their research.

On the one hand, individual responses to mask mandates, lockdowns, school closures, hospital overflows, toilet paper shortages, and vaccine provision generate a vast range of reactions from different segments of the population. Some people have no problem wearing a mask but find school closures unbearable. Some find the thought of vaccine passports to be an autocratic imposition on their freedoms, even after willingly vaccinating against other diseases. Some think that wearing a mask should be a personal choice but take offense at others who make this choice themselves. As these populations appear unwilling to comply with expectations, their behavioral changes serve to prevent or thwart the successful implementation of such approaches. Meanwhile, their anticipated outcomes are based on model assessments of how aggregations of individuals behave in the face of health, economic, and social considerations.

On the other hand, state responses to our COVID-shaped social reality face a similarly broad set of reactions. Lockdown decisions by municipalities and states are met with complaints of governmental or institutional overreach, as well as complaints of insufficient action. Mask mandates are found by some to be an outrageous restriction on individual freedoms; others find them to be a toothless bit of "pandemic theater." School closures are thought by some to compromise opportunities to which schoolchildren have a right, while others see those closures as a minimal means of protecting the well-being of those same children. As municipalities and states attempt to develop policy responses to the rapidly evolving COVID-environment, they make predictions based on model assessments of how this or that policy option will be received by the general population.

Both individual and state responses might be understood in risk assessment terms. For instance, individual or collective choices might be modeled as inputs that determinately shape the output state of affairs: given some probability of an anticipated outcome, Φ-ing will promote or prevent that outcome. But there is an important sense in which this characterization in risk assessment terms underdetermines the complex problems that often inform such responses. Indeed, subsequent decisions by individual agents made in light of the complexities associated with COVID change the risk assessment informing future decisions, thus transforming ostensibly determinate systems into indeterminate systems. This problem is amplified in emergent phenomena like COVID, where disease prevalence and human behavior are ever-changing, partly as a consequence of the risk modeling that is done to inform policy. When a decision to open schools or restaurants is based on current risk assessments, for instance, these risk assessments become obsolete the moment that they are translated into action. This is the *problem of indeterminacy* that interests us here. In what follows we hope to explain the extent to which indeterminacy is pervasive throughout the policy-making apparatus, explain the way in which naked risk assessments taken out of context can upend our best intentions, and diagnose those upended intentions.

3.3 Background

3.3.1 Responsibility and Causality

There is, conventionally, a close relationship between responsibility and risk. In traditional models of responsibility, a causal connection between an action and an outcome is generally assumed. This causal supposition serves as a foundation for related attributions of blame (through moral responsibility), liability (through legal responsibility), and a good deal of our moral and normative conceptual apparatus.

The idea that agents are accountable for actions they perform, resulting in predictable states of affairs, is commonplace. For instance, if Stearns throws and breaks a ceramic pot, so long as he recognizes that the pot had a reasonable risk of breaking, he is responsible for the broken pot. Whether Stearns proceeds to throw the pot involves a bare-bones and rudimentary risk assessment. He must ask himself: what is the probability that the pot will break? His assessment of the outcome, shaped by his understanding of the risk of the outcome following from his action, shapes his understanding of his own responsibility both before and after the event.

Of course, there are many other factors that influence our thinking about the outcomes of such behavior. Was there an external force that compelled Stearns to throw the pot? Was the pot cracked in advance of his decision to throw it? Were there environmental features that compromised the integrity of the pot? These factors are typically spelled out in terms of risk assessments: decisions, actions, and outcomes that account for outside parties or forces. Responsibility is thereby often treated as a spidery network of attributions tracing back to the sundry actors that are commingled in the eventuating outcomes.

Our focus in this chapter is less on moral responsibility and more on the deterministic framework that undergirds the conventional approach to causal responsibility. The attendant moral theories that link moral responsibility to causation are obviously affected by the position we argue for here inasmuch as they are dependent upon views about causation, but they are not the central focus of our argument and lie outside the scope of this chapter.

The risk model that interests us, along with its attendant framework for understanding responsibility, is markedly individualistic. It relies on isolated individual actors bringing about states of affairs deterministically, as if they are cogs in a machine, marbles in a marble run, drugs in a bloodstream. While notions of complicity and aggregation provide some level of nuance to the individualistic account, the foundation of responsibility remains deterministic.

Expected utility theory and its assumed "rational actor model" are closely tied to this individualistic and deterministic account of responsibility (Morgenstern and Von Neumann 1953). Generally understood, the rational actor model construes agents as decision makers who choose between outcomes based on interest optimization or maximization. On this account, options, understood in terms of outcomes or states of affairs, are characteristically ranked according to desirability or preference. The rational actor model thereby serves as a central framing mechanism for risk analysis and any attempt to anticipate or manage risk. But we will argue below that restricting our understanding of responsibility and human behavior to a risk-management approach is dangerously misguided.

Three separate observations may be helpful here in establishing our position, each covered in the subsequent three sections. (A) The "risk society" is all-encompassing and manufactured risks are prevalent. (B) The colonization of our lifeworld and the related idea of strategic thinking have shaped and limited our understanding of possible responses to wicked problems. (C) The indeterminacy of the real world, coupled with strategic thinking, points to the inevitable mismanagement of those problems.

3.3.2 The Risk Society: Manufactured Risk

In the 1980s and 1990s, sociologists Ulrich Beck and Anthony Giddens introduced the notion of a "risk society" to the academic community (Beck 1995, 2012, 2018; Giddens 1984, 1990, 1999). Their concerns take their stepping off point from the characterization of risk and responsibility that we mention above, though they go one step further in teasing out the implications of risk management. Essentially, Beck and Giddens acknowledge that while human civilization has always been subject to hazards that threaten to upend our stable lives – threats to life, limb, happiness – it is through the conceptual innovations of modernity – among other things, the risk-cost-benefit decision trees we just mentioned – and the imposition of technical interventions aimed to mitigate these risks, that we have learned to approach and manage risk differently than ever before. At the same time that we succeed in mitigating natural risks, the very risk management technologies we implement to mitigate those risks themselves become reflexive, generating a host of new, sometimes more existentially dangerous, risks.

Perhaps the simplest and most concrete example of this comes from the nuclear power industry, which is what concerned the two theorists when they were first writing. In building an industry aimed at addressing energy insecurity – by constructing nuclear power plants – we have peppered the landscape with nuclear plants that risk melting down. The solution to one risky scenario introduces new, in this case arguably more challenging, risks. Chernobyl, Three Mile Island, Fukushima Daiichi all serve as reminders that the risks in our energy production sector are real. The 2022 Russian invasion of Ukraine and the related assault on the Chernobyl disaster area (in which Russian soldiers held technicians hostage and failed to recognize risks to themselves and the country) serves as a stark reminder of just these sorts of introduced or "manufactured" risks that would not be in place but for the risk mitigation efforts in the first place (Kramer 2022). Much the same sort of argument can be made for other attempts to mitigate risk as well: the Deepwater Horizon oil spill, the setting aflame of derricks during the Persian Gulf War, the Union Carbide disaster in Bhopal, and of course the 20 year US occupation of Afghanistan.

It is through this discussion that Beck and Giddens aim to distinguish between "external risks" and "manufactured risks," the latter of which being of central concern to them. In its simplest form, the thought is that external risks are posed by the universe, whereas manufactured risks are imposed by our attempts to mitigate risks. It would be easy to look at the superficial dimensions of Beck's and Gidden's work and assume that it is only ever the technologies themselves that introduce manufactured risks. But this is too narrow. It is the very calculation of risk that gives rise to the risk society, to a society in which risk is the defining heuristic through which to approach the future.

3.3.3 Colonization of the Lifeworld: Strategic Thinking

Pair this discussion of manufactured risk with some of the observations made by philosopher Jürgen Habermas, who has written extensively on a phenomenon that he calls "The Colonization of the Lifeworld." In much of his early work, Habermas was troubled that the conceptual innovations of modernity themselves, not just the technologies that have been put in place to manage our risks, pose yet a different kind of concern (Habermas 1972, 1987b, a, McCarthy 1978). Though the technical nuances of this colonization are in a way immaterial to the upshot of this paper, Habermas borrows the notion of the "lifeworld" from Edmund Husserl (1970) and the notion of the "system" from Talcott Parsons (1951/1991). He brings these two ideas together to advance a "two-level social theory," suggesting that there is an important sense in which over the course of our lives, we live and operate in the social realm on two levels at once. On the one hand, we live among humans, engaging in the daily activities of the lifeworld that are defined and governed by cultural conventions and interpersonal norms. On the other hand, we live in a world that has been systemized – made sense of, modeled, and manipulated by techniques of rational conceptualization.

For Habermas, the lifeworld in question is fairly abstract: a series of socially and culturally sedimented conventions and norms that make up the background experiences of each one of us. Depending on the social and cultural framework in which we come to understand the world in which we live, we tend to approach the world and the challenges of the world, in ways that reflect these socially and culturally sedimented norms. Colonizing the lifeworld amounts to rationalizing and smoothing out the rough patches – the difficult bits, the complex parts – that may be culturally or conventionally informed. If we carry this back to the way in which we approach risk and responsibility, this might mean, for instance, that where we once approached natural hazards fairly directly – preventing harm from befalling us by wearing armor, brandishing a sword, building a bridge – risk assessment has allowed us to approach natural hazards and manufactured risks as the sorts of challenges that can be made sense of in actuarial terms. In systematizing the world around us – that is, modeling the world and putting it into a context that we can understand rationally – we effectively transform the way we understand and interact with the chaotic world we inhabit most of our days.

Such systems are developed with the express purpose of making sense of our world so that we can control it. In turn, systems-thinking is tied to Habermas's notion of strategic action, where all actions are taken with the objective of achieving a desired end, not necessarily with the mutually supportive objective of arriving at an understanding. By assuming a strategic stance in approaching problems of risk, we impart to our scenarios a distortionary perspective that suggests that a solution cannot be arrived at through any means other than by taking action unilaterally, whether at the individual or the state level. In turn, this stance *precludes*, as a conceptual matter, an alternative approach to coordinating action – one that has both historical roots and future potential. Defaulting to strategic action constrains our worldview such that we tend to undermine our own capacity for judgment about how to arrive at a mutually acceptable and cooperative agreement.

In this way, Habermas argues that the colonization of the lifeworld, manifesting often as systems of economic or bureaucratic imperatives, displaces the normative capacity of language and, in so doing, undercuts the emancipatory potential of ethical or democratic engagement.

3.3.4 The Self-Undermining Nature of Risk Assessment: Indeterminacy

To say that a model is determinate is just to suggest that the model anticipates that the universe will turn out in a particular kind of way given whatever inputs push it that way. How marbles will spill out of a jar onto the floor, which direction projectiles will fly if launched into a headwind, how satellites travel around the globe, whether a trolley on a track at a given speed will strike five bystanders – are all cases of determinate systems. Indeed, it turns out that many decision trees conceive of the key decisions they seek to model in determinate terms, even if the outcomes are anticipated across wide, probabilistic error bars.

But models of future outcomes are complicated by the presence of intervening agents, particularly if those agents are strategically motivated (Hale 2022; Hardin 2003). Many cases of anticipated outcomes involve human agents, all of whom maintain the capacity to respond to scenarios as they change. These agents not only respond to scenarios as they change but also change those scenarios such that the models become obsolete the moment they are translated into action by intervening agents.

The COVID cases we mention above offer concrete examples of this, and yet also show how limiting such a way of thinking can be. Take just one familiar example: a common approach to deciding when it is safe to re-open schools to in-person learning has been to look at disease prevalence (Miller, Sanger-Katz, and Quealy 2021). But disease prevalence is, of course, determined by the behavior of humans. If disease prevalence reports are used in risk assessments that determine what policy will govern school openings, or whether an individual agent is comfortable sending her kids to school, this will in turn change the dynamics of the model by ultimately affecting disease prevalence too. In this way, COVID prevalence is not a stable, but rather an emergent, phenomenon.

The determinate risk models used to assess the spread of COVID have been regularly undermined in this way, no matter whether the solution uses the individual or the state as the primary unit of analysis. It is by now routine to encounter robust models and arguments that appear a few weeks prior to anticipate one set of outcomes only later to be shown false. What is interesting is not *that* they are undermined, but *how* they are undermined. And what is particularly interesting is that they are *self-undermining* – in part because they conceive of the world in these deterministic terms – as if individuals and states are compelled by social laws in the same way that physical objects are compelled by laws of nature.

Policy-makers often attempt to shoehorn the rubrics of determinate modeling into decision-making, as they desperately try to fashion responses to the allegedly uncooperative behavior of individuals. Say, for instance, as they respond to people who won't get vaccinated, who refuse to wear masks, or who attend random bike rallies in North Dakota. In all of these cases, we see the same pattern. Assessments of these problems then characteristically become *reflexive*, with modelers returning to the outcomes that they have anticipated, puzzling over the fact that their anticipated outcomes have been subverted and not come to pass, all the while attributing to actors and policy-makers a logic that explains how the observed outcome did in fact come to pass. They might, for instance, explain the proliferation of disease after aggressive vaccination regimes as a kind of tragedy of the commons: that individual actors are acting selfishly and in so doing putting the rest of the community further at risk.

3.4 The COVID Commons

One of the more egregious and common instances of such a reliance on modeled outcomes occurs with Garrett Hardin's classic Tragedy of the Commons. In that work, and in the voluminous literature that follows from it, Hardin suggests that self-interest and scarcity are sufficient to explain how a commonly shared resource – a "commons" – comes to be degraded (Hardin 1968). He implies, in doing so, that the solution to the problem is to change the incentive structures. The prevailing assumption here is that there is an undesirable outcome – the presumed "tragedy" – and that our creation of this tragedy can be made sense of by looking at the various strategic pressures on any or all participating parties.

As it happens, there is some reason to doubt that Hardin's so-called tragedy is as clear as it initially appears. To explain, we can isolate three perfectly reasonable interpretations of what makes the tragedy of the commons so tragic. The first might be that it degrades values, the second that it is self-undermining, and the third that the future appears to be "locked in" by dint of a particular strategic orientation, a logic of strategy. On the first interpretation: the most direct and superficial observation about the tragedy of the commons is that the commons is degraded and that this degradation is what makes the tragedy tragic. When lakes or fields or atmospheres are so degraded that they lose value, all thanks to the behaviors of hundreds, thousands, or even billions of overlapping actors, this is a tragedy that our scions of industry and producers of value ought not to countenance. It is a tragic *loss to the universe*.

On the second interpretation, a slightly less direct pathway to understanding the tragedy of the commons is to view the tragedy as one of undermining one's own objectives in the name of pursuing one's own objectives; undermining their aims in the way that fishermen who destroy their fishery might be undermining their very livelihoods. The thought here is that each actor takes actions to pursue her own interest but self-undermines this objective and ends up in a suboptimal state regarding her own interest. Hardin worries that ruin is the destination toward which all men rush, presumably suggesting that the tragedy rests in the self-undermining nature of the tragedy of the commons. This is a tragic case of *shooting oneself in the foot*.

On the third interpretation, the sense in which the tragedy of the commons might be tragic comes in its apparent inevitability. Inasmuch as there appears to be no viable solution to the tragedy of the commons, in that all dominant strategies point to a highly suboptimal outcome, there's almost nothing that can be done to stop this. As we (inevitably) pursue our own ends, without the politically infeasible heavy-handed regulation of the state or the pointless self-sacrifice of individuals, we will find ourselves, however well-intentioned, destroying our commons. The tragedy of the commons is tragic because it is *an impossible wicked problem*. But, of course, it is not so inevitable. We can and have overcome the tragedy, and so it is not and should not be thought of as some law of nature.

We think the above discussion of indeterminacy reveals yet another way to understand what's so tragic about the tragedy of the commons. What the tragedy of the commons reveals when one properly understands the indeterminacy that it engenders is that the tragic failure to cooperate arises not because communication cannot happen, but because *all parties are strategically inclined*. They simply neglect to approach the question of preferable outcomes from the standpoint of communicative engagement because they instead approach it as an interest maximization problem.

It is easy to miss this bit about strategy because the tragedy of the commons as typically modeled directs our gaze toward the suboptimal outcome – the sad little box in the lower right-hand corner of the payoff matrix – and this assumes that cooperation is synonymous with communication. But the tragedy of the commons arises in games of pure coordination where, though the interests of actors are in cooperating to achieve a collective outcome, the actors are nevertheless doomed to fail because they cannot communicate. Since the game is so structured, it gives the *appearance* that the central problem with the tragedy of the commons is that participants are not able to arrive at an optimal outcome, again, largely because they cannot coordinate (which in this case is the presumptive form and purpose of communication). The solution, as a consequence, is to get them to "communicate" (i.e., coordinate) in some way. Typically, this is done by aligning payoff structures or by encouraging control. Solutions to our highly and artificially constrained individual deliberators are framed in terms of collective action. In this framing, once we have convinced ourselves that there are no viable individual solutions, we are left with only one choice: solve the "tragedy" through collective coercion.

But it is worth reiterating that this is not the true tragedy of the tragedy of the commons. The tragedy is that people are approaching one another *strategically*, assuming that there are other barriers to cooperation, and resisting communication. They are treating one another as mere objects of causal significance. And that returns us to the point of responsibility attribution we briefly mention at the beginning of this chapter.

Many pictures of responsibility are deflationary, in the sense of reducing the complexity of the real world to what is perceived as the essence of some problem. Of course, the worry with this deflation is that the model it uses to describe the world leaves things out. We see this in the philosophical literature on promising, where the language of promising is reduced to expectations (Mason 2005). We also see it in the literature on harm (Bradley 2012), where the rich language of harm is reduced to the undermining of interests (Feinberg 1987). There may well be good reasons to reduce and apply models that simplify explanation, in cases like promising – where the language of promising involves appealing to, say, respect and status (Darwall 2006; Gilbert 2011; Shockley 2008) – or harm – where the language of harm captures a sense of wrongdoing beyond the undermining of interests (Feit 2015; Shiffrin 2012), but there are costs to this sort of deflation. In accepting an account of responsibility reduced to causal relationships, one risks treating individuals as mere objects of causal significance. The framing of individual responsibility as bound up in determinate causal networks, while useful to risk analysis, deflates the complex ways in which responsibility might fit into our analysis of human behavior. And it thereby restricts the options available to us when we try to assess responsibility in the real world (Young 2011).

Similarly, Chris Cuomo worries that common framings of responsibility for climate change rely on either individual or state-based responsibility, and so frame the problems of climate change in ways that obfuscate the historical and structural pressures that led to the current problem. Such framings render the real cause invisible. Cuomo suggests we focus on "metalevel responsibilities" (Cuomo 2011, 704), responsibilities of entities like utility companies, corporations, and non-state actors, and we agree (Boran and Shockley 2021; Hale 2020). The individual-state dichotomy is a deflation of the more complex social milieu in which we deliberate and on the basis of which we shape our analysis. The reliance on an exclusive and exhaustive individual-or-collective framing of responsibility constitutes a paradigmatic example of how a model colonizes our analysis and thereby artificially constrains our range of options.

3.5 Decolonizing the Lifeworld: Reinflating Our Options and Possibilities

Let our responsibilities track a more realistic social ontology. What we hope to have suggested above is that it is a feature of our talk of risks and decision-making that we treat individuals as determinate entities, not agents or actors. This is a two-pronged simplification. On the one hand, it helps make manageable our deliberations about what to do: we artificially limit the expected outcomes to those we can anticipate, all else being equal. On the other hand, the cost of this simplified and manageable model is to ignore the complex ways that humans as agents respond to social pressures and information, including pressures and information that themselves are predicated on the simplified model. Human agency introduces considerable indeterminacy that lays bare problems with the determinate worldview at the heart of the risk society (Hale 2022; Hardin 2003).

The important epistemic observation we make here is that even though risk assessments may serve as helpful heuristics for making sense of dynamic systems, they cannot serve as instruction manuals on how to tweak those systems to achieve a desired outcome. Recognizing instead the role of human agents in the creation of outcomes will mean recognizing in turn the fundamentally indeterminate nature of our present world, hopefully adding a little bit of nuance to an otherwise desiccated picture of social problems.

3.6 A Third Way

Sometimes our problems do not require anticipating the future and optimizing among predetermined options. In these indeterminate cases – COVID, climate change – it is an error of practical reason to assume that such cases can be addressed in this way. What we need instead is a process with considerably more agility in order to respond to circumstances as they arise. We need to integrate adaptation and agency not only into our deliberative process, but also into the way we frame complex problems.

In everyday interactions, we typically don't engage in this sort of risk assessment. Think of Darwall's "Two Kinds of Respect" (1977) – we shouldn't engage with others as if they are objects to be manipulated, normally at least. To do so is to fail to treat them as persons. To engage with them, normally, is to presume they are subject to and grappling with reasons in the same way as all others (c.f., Pettit and Smith's appeal to the Conversational Stance [1996]). The difference here rests in two kinds of respect: between respecting your opponent and respecting your "opponent's left hook" (Darwall 1977, 40). While both may be warranted, failing to respect your opponent as a person is a practical failure … a failure of practical reason. To understand a boxing match as more than a description of strikes and dodges, one should focus on the boxers as much as their actions. The agency of the boxer matters. The agency adds indeterminacy and makes strategic engagement inappropriate (or at least misguided). Strategic engagement is different from communicative interaction. The point that we are making is less that these kinds of orientations fail to respect other persons, though that may be true, and more that it is a mistake of practical reason to approach indeterminate problems as just very complex risk problems. Doing so assumes that social hazards can be addressed in the same way that we might go about addressing hazards from catastrophic volcanic eruptions or asteroid strikes.

Now then, this is not to say that there is no sense in using risk models to help understand how the future might unfold, but *understanding* or *anticipating* future outcomes alone is not the upshot of many risk management regimes. Risk analyses are built around a specific purpose: namely, they aim to *alter* the risk, and they seek to do so by using the selfsame levers that are incorporated into the risk models themselves. It is the mistaken assumption that we are all strategic agents who act merely as levers of change that creates the complication. In both state-based solution orientations – where states are thought to coordinate action through coercive mandates and incentives – and individual-based solution orientations – where aggregations of individuals are presumed to act harmoniously according to the same (determinate causal) logic – the mistake is to assume that there will not be actors who aggressively aim to subvert the solution.

Elinor Ostrom's work in New Institutional Economics points to the possibility of there being an alternative to the determinate model of human action and the use of the individual or the collective as the primary unit of analysis (Frischmann 2013; Ostrom 1990, 2000). Ostrom's work shows in part how the tragedy of the commons is an artifact of a mistaken determinate model of human behavior. In order to arrive at the tragedy in the first place, risk analysts are invited to model people as rational strategic actors who can be manipulated as objects, pushed, and prodded through causal pressure such that the optimal solution is found by prodding them in the right way, thereby compelling optimal behavior.

It is perhaps worth noting that, particularly given the poor track record of regulating the commons, if the logic of the tragedy of the commons were right, civilization should have already been driven to ruin, quickly and efficiently. Says Hardin: "Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons. Freedom in a commons brings ruin to all" (Hardin 1968). Indeed, all of our commons areas should by now already be despoiled. While it is true, of course, that many commons areas are in fact under pressure – from fisheries to atmospheric carbon to potable water – they remain not as ruined as Hardin's analysis would have us believe.

The significance of the social world, often reduced or minimized in rational actor models, illustrates the failing of approaching future outcomes too deterministically. Among other things, this is partly what Ostrom's research reveals: that in many smaller group contexts people find ways of regulating their behavior through means other than deterministic individual constraint or political regulation. What Ostrom's work leaves off the table, however, is the mechanism by which norms take root, whether they be norms of behavior *between* individuals within collectives or between the individual and the whole collective. We have little sense from her empirical research of how these norms propagate through communities, why they take root the way they do. And the worry here is that without an explanation for this, we may persist in our thinking that norms don't take root, they just operate as if by magic. What we are offering in this paper, instead, borrowing heavily from the work of other philosophers like Stephen Darwall, Jürgen Habermas, Rainer Forst, Philip Pettit, John Dryzek, and of course many others, is the idea that these norms take root and steer collectives of individuals through discursive engagement, and that in virtue of this, risk *modeling* cannot translate, and ought not to be presumed to translate, into risk *management* directly (Darwall 2006; Dryzek 2000; Forst 2012; Habermas 1987a; Pettit 2012).

Taking a risk management perspective, with all of its strategic rationality, is itself a manufactured risk. The reliance on such a perspective involves, then, a form of *mis*management. As Cuomo (2011) notes, the approach to environmental problems that is framed in terms of individual obligation sells us down the river. It prevents putting leverage on legal, economic, and political structures. By addressing the ethics of climate change through a framework limited to individual obligation, we fail to recognize other possible sources of responsibility and, in so doing, disregard other possible levers for making positive and productive changes in our responses to a changing climate.

More is needed than individual action (it is often said), but with a framework limited to individual obligation, we are blind to many of those options (Young 2011). An individualistic framing of responsibility hampers our ability to respond effectively to climate change. In responding to challenges such as climate change, COVID, or fisheries depletion, it is not necessarily the case that we choose not to endorse or accept some alternative framework. Rather, the challenge is that our approach to these problems is limited by the way our lifeworld has been colonized by this specific way of understanding responsibility. Our own take is that solutions should be sought through institutional arrangements tied up with deliberative democracy.

COVID is a cautionary tale about how individuals might be manipulated, strategically, in support of a set of ostensibly predetermined ends. With the political landscape of COVID framed as a grand conflict between individual freedom and the public good, with outcomes framed in terms of risk management, the set of options available and the set of outcomes that might occur are limited. But the strategic risk management approach cannot sufficiently capture the indeterminate effect of individuals operating as agents non-strategically.

Consider again the release of a COVID vaccine. We can try to model the behavior of those who don't want to take the vaccine as the vaccine hesitant, but those who resist vaccines bring to the table a bunch of different and sometimes conflicting reasons for their hesitancy. Maybe they're concerned about efficacy. Maybe they're concerned about safety. Maybe they're concerned about the encroachment of their personal liberties. Maybe they're under the spell of a persuasive and charismatic orange idol. Maybe they're reflecting a long sordid history of manipulation by government in their community. The list of various reasons that any individual, not to mention group of individuals, has for resisting vaccination is long … certainly much longer than we can articulate here.

Again, we can try to model these reasons in different demographic groupings, maybe even parceling them out according to their approximate percentages. Our pollsters and pundits have made a cottage industry of this way of thinking. But doing so invariably leaves unexamined the variety of ambiguities that exist within these groupings, that understands these groups, and the deployment of the reasons that motivate them, always through the lens of those doing the modeling.

Moreover, when individuals, acting on whatever reasons they might have, interact with other individuals, acting on whatever reasons that their group might have, the results are both highly unpredictable and subject to revision with the introduction of each new reason. As Jonathan Dancy might say, "reasons are like rats" (Dancy 2004, 15): even if you can understand one reason, you may know little to nothing about how it interacts or overlaps with other reasons. The same might be said of people, and their reasons, particularly when under the stresses predictably associated with COVID and other life altering features of this, our modern age. We would do well not to presume that people as individuals, nor people in groups, will act according to a preset selection of options or reasons, but rather must recognize individuals as engaged in an exchange of reasons with others. Their reasons – behaviors, actions, maybe even their beliefs and desires – change with pressures from within or without.

And this problem of indeterminacy in our risk models is far from limited to global pandemics. We haven't been able to anticipate the vicissitudes of the stock market, the vagaries of the political scene, how populations respond to natural disasters, where factions will rise or fall in war, and so on and so forth. We don't know how people are going to respond to various interventions or stressors because the reasons they have for taking the actions they take are not on view for all to see. Moreover, those reasons shift and change with the shifting and changing landscape in which they are operating. Consider the 2016 US presidential election. In predicting Hillary to win, people respond to such news in different ways. Individuals shift their motivations. Reasons and actions change in unpredictable and indeterminate ways, both to new information and to the behaviors and actions of others.

Our point isn't so much that we will be able to pull ourselves out of the COVID crisis, or any other crisis of our current age, by becoming more democratic – essentially holding a bunch of focus groups and citizen juries. Unfortunately, we've already set the table for parties to engage oppositionally. Many of the political problems we face now have come about because we have grown dependent upon the machinations of public health policy construed as risk management. We don't have a suitable democratic apparatus in place to resolve these political conflicts because we have, historically and over the development of the Western model of democracy, repeatedly prioritized a picture of the future that fails to account for the agency of citizens to respond in their own ways, whether individually or collaboratively. And there are really two problems here. On one hand, we are never going to get the internal point of view that provides the basis of reasons upon which people act if we stick with the present risk-oriented approach to the future. Individual perspectives will never be adequately captured by modeling. On the other hand, people behave and act differently when faced with reasons, behaviors, and actions of others. When the exogenous forces of other humans or other pressures come into play, this makes modeling incredibly difficult.

What we need to do instead, in anticipation of impending crises – either crises that are slowly unfolding like climate change or rapidly emergent like the next pandemic – is to create the conditions for democratic and discursive engagement. This means changing rules about how information is disseminated so that, say, it is not funded and directed by interested parties that aim to sell something. It means encouraging reflection in schools and turning away from rote regurgitation. It means creating spaces for members of the public from differing backgrounds to meet and engage one another. It means providing a sufficient economic baseline for them to be able to effectuate policies in play.

The idea, again, is not that a golden democracy on the hill is likely to come to pass, nor that a turn toward discursive democracy will solve our COVID crisis overnight, nor even that models are useless in helping us understand pressures that may be in play, but rather that risk models, when applied to populations as management guidebooks, are *bound to fail.* They are bound to fail because they assume determinacy as one of their central suppositions.

3.7 Conclusion

The attempt to understand the future in terms of risk – with hazards and probabilities that are epistemically evaluable through modeling and projection – is undercut by observations from social choice and game theory: the idea that knowledge about the future is uncertain, but that the future itself is heavily dependent upon actions taken by actors in the present. This manifests in so many ways that go beyond what we've been able to cover here: the IPCC's five different pathways for climate impacts; trading pork futures on the stock market; tranches and junk bonds; subprime mortgages; guerilla warfare; terrorism; and on and on and on.

Yet, for centuries, humans have been finding ways of coordinating action, among multiple parties, without assuming that the correct lens to coordinate that action is risk analysis. One might think that the reaction to the tragedy of the commons should not be how we might avoid wrecking the commons, but why we haven't done so already. If the tragedy of the commons is correct, the destruction might well be inevitable – nearly a law of nature. Yet we find ways to coordinate, without heavy handed regulation, without self-sacrifice. We collaborate. And we've done so for as long as we've shared resources or shared a lifeworld.

In the face of the collective risks of COVID, climate change, fisheries exploitation, pollution, etc., classical liberal ethicists have devolved power and responsibility down to the individual, effectively neglecting centuries of efforts to coordinate action. This move has resulted in a false choice between individual and collective approaches, which minimizes or ignores the complex realities of our social environment. Above, we address the responsibilities of dealing with climate change and COVID by considering the ways in which this individual-state approach to responsibility has colonized our lifeworld and restricted the range of possible responses to such problems. Acknowledging the indeterminacy implicit in so-called Wicked Problems (Lazarus 2009) – whether in response to pandemics, addressing climate change, or dealing with the threat of tragedies of the commons – should push us toward a different response. Further models relying on the presumptively determinate nature of human behavior will predictively fail. A deliberative approach, one that embraces our indeterminate future, seems a better option.

Note

[1] On January 30, 2020, the World Health Organization (WHO) declared the global outbreak of COVID-19 to be a public health emergency of international concern (PHEIC). On May 5, 2023, the WHO declared an end to COVID-19 as a PHEIC.

References

Beck, Ulrich. 1995. *Ecological Politics in an Age of Risk*. Cambridge: Polity Press.

Beck, Ulrich. 2012. "World Risk Society." In *A Companion to the Philosophy of Technology*, edited by Jan Kyrre Berg Olsen Friis, Stig Andur Pedersen and Vincent F. Hendricks. London: Wiley-Blackwell.

Beck, Ulrich. 2018. *The Reinvention of Politics: Rethinking Modernity in the Global Social Order*. Cambridge: Polity Press.

Boran, Idil, and Kenneth Shockley. 2021. "Governance Toward Goals." In *Principles of Justice and Real-World Climate Politics*, edited by Sarah Kenehan and Corey Katz. Lanham: Rowman & Littlefield.

Bradley, Ben. 2012. "Doing Away With Harm." *Philosophy and Phenomenological Research* 85 (2): 390–412. https://doi.org/10.1111/j.1933-1592.2012.00615.x

Bramble, Ben. 2020. *Pandemic Ethics: 8 Big Questions of COVID-19*. Sydney: Bartleby Books.

Cuomo, Chris J. 2011. "Climate Change, Vulnerability, and Responsibility." *Hypatia* 26 (4): 690–714. https://doi.org/10.1111/j.1527-2001.2011.01220.x

Dancy, Jonathan. 2004. *Ethics Without Principles*. Oxford: Clarendon Press.

Darwall, Stephen. 1977. "Two Kinds of Respect." *Ethics* 88 (1): 36–49.

Darwall, Stephen. 2006. *The Second Person Standpoint: Morality, Respect, and Accountability*. Cambridge: Harvard University Press.

Dryzek, John S. 2000. *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford: Oxford University Press.

Feinberg, Joel. 1987. *Harm to Others*. Oxford: Oxford University Press.

Feit, Neil. 2015. "Plural Harm." *Philosophy and Phenomenological Research* 90 (2): 361–88.

Forst, Rainer. 2012. *The Right to Justification: Elements of a Constructivist Theory of Justice*: Translated by Jeffrey Flynn. New York: Columbia University Press.

Frischmann, Brett M. 2013. "Two Enduring Lessons from Elinor Ostrom." *Journal of Institutional Economics* 9 (4): 387–406. https://doi.org/10.1017/S1744137413000106

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley: University of California Press.

Giddens, Anthony. 1990. *The Consequences of Modernity*. Stanford: Stanford University Press.

Giddens, Anthony. 1999. “Risk and Responsibility.” *The Modern Law Review* 62: 1–10.

Gilbert, Margaret. 2011. “Three Dogmas about Promising.” In *Promises and Agreements: Philosophical Essays*, edited by Hanoch Sheinman, 80–108. Oxford: Oxford University Press.

Habermas, Jürgen. 1972. *Knowledge and Human Interests*. Boston: Beacon Press.

Habermas, Jürgen. 1987a. *The Theory of Communicative Action*, *Volume 1: Reason and the Rationalization of Society*. Translated by Thomas McCarthy. Vol. 1. Boston: Beacon Press.

Habermas, Jürgen. 1987b. *The Theory of Communicative Action*, *Volume 2: Lifeworld and System*. Translated by Thomas McCarthy. Vol. 2. Boston: Beacon Press.

Hale, Benjamin. 2019. Paper presented at the University of Reading, the University of New South Wales, the University of Colorado, Boulder.

Hale, Benjamin. 2020. “Right-Leveling Indeterminacy: Environmental Problems, Non-State Actors, and the Global Market.” In *Climate Justice and Non-State Actors: Corporations, Regions, Cities, and Individuals*, edited by Jeremy Moss and Lachlan Umbers. New York: Routledge.

Hale, Benjamin. 2022. “Indeterminacy and Impotence.” *Synthese* 200 (250): 1–24. https://doi.org/10.1007/s11229-022-03718-7

Hardin, Garett. 1968. “The Tragedy of the Commons.” *Science* 162: 1243–48.

Hardin, Russell. 2003. *Indeterminacy and Society*. Princeton: Princeton University Press.

Husserl, Edmund. 1970. *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*. Evanston: Northwestern University Press.

Jaziri, R., and S. Alnahdi. 2020. “Choosing Which COVID-19 Patient to Save? The Ethical Triage and Rationing Dilemma.” *Ethics, Medicine and Public Health* 15: 100570. https://doi.org/10.1016/j.jemep.2020.100570

Kramer, Andrew E. 2022. “Chernobyl as Staging Ground? Russians Ignored Warnings.” *New York Times*, April 9, 2022, A.

Lazarus, Richard J. 2009. “Super Wicked Problems and Climate Change: Restraining the Planet to Liberate the Future.” *Cornell Law Review* 94: 1153–234.

Mason, Elinor. 2005. “We Make No Promises.” *Philosophical Studies* 123 (1–2): 33–46.

McCarthy, Thomas. 1978. *The Critical Theory of Jürgen Habermas*. Cambridge: MIT Press.

Miller, Claire Cain, Margot Sanger-Katz, and Kevin Quealy. 2021. “Schools Don’t Need Vaccines to Hold Classes in Person Safely, Experts Broadly Say.” *New York Times*, Feb 12, 2021, A.

Morgenstern, Oskar, and John Von Neumann. 1953. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. London: Cambridge University Press.

Ostrom, Elinor. 2000. “Collective Action and the Evolution of Social Norms.” *Journal of Economic Perspectives* 14 (3): 137–58.

Parsons, Talcott. 1951/1991. *The Social System*. London: Routledge.

Pettit, Philip. 2012. *On the People's Terms: a Republican Theory and Model of Democracy*. Cambridge: Cambridge University Press.

Pettit, Philip, and Michael Smith. 1996. "Freedom in Belief and Desire." *Journal of Philosophy* 93 (9): 429–49.

Reddy, Sanjay G. 2020. "Population Health, Economics and Ethics in the Age of COVID-19." *BMJ Global Health 5* (7):e003259. https://doi.org/10.1136/bmjgh-2020-003259

Rodger, Daniel, and Bruce P. Blackshaw. 2022. "COVID-19 Vaccination Should Not Be Mandatory for Health and Social Care Workers." *The New Bioethics* 28 (1): 27–39.

Shiffrin, Seana Valentine. 2012. "Harm and Its Moral Significance." *Legal Theory* 18(3): 357–98. https://doi.org/10.1017/S1352325212000080

Shockley, Kenneth. 2008. "On that Peculiar Practice of Promising." *Philosophical Studies* 140 (3):385–99. https://doi.org/10.1007/s11098-007-9151-7

Varoufakis, Yanis. 2013. *Economic Indeterminacy: A Personal Encounter with the Economists' Peculiar Nemesis*. London: Routledge.

Young, Iris Marion. 2011. *Responsibility for Justice*. New York: Oxford University Press.

Part II

Legal Context

4 Risk, Responsibility, and Pre-Trial Detention

RA Duff

4.1 Introduction: Pre-Trial Detention and the Presumption of Innocence

When we think about our prisons, we might think of them as populated by convicted offenders serving terms of imprisonment. However, a substantial number of prisoners are on remand awaiting trial: 8,304 (10.5%) of the total population of 79,092 in English prisons, for instance, on December 31, 2021;[1], 2,002 (26.7%) out of a total population of 7,502 in Scottish prisons at February 4, 2022.[2] For many of them, their detention is relatively short; in Scotland, for instance, the median time on pre-trial remand in 2019–20 was 21 days.[3] But for some it exceeds a year: in England, at December 31, 2021, 4,185 had been awaiting trial for over six months, of whom 1,710 had been detained for more than a year, 480 of them for more than two years.[4] Many in pre-trial detention are then convicted and sentenced to imprisonment, but many are not: in England, about 15% are acquitted, and about 30% receive a non-custodial sentence after conviction.[5]

Pre-trial detention is frequently harsh and damaging. Prison conditions are often, at best, austere; remand prisoners do not have access to the kinds of training, education, or work that are available to sentenced prisoners; they may find it hard to consult their lawyers, or to keep in touch with their families; their relationships are likely to be damaged; they might lose their jobs or their homes. But they have not been proven guilty of the crimes for which they face trial: the courts must, supposedly, presume them to be innocent. How could such detention be justified, given its apparent inconsistency with the presumption of innocence and the right to liberty enshrined in the European Convention on Human Rights?[6] Why does this practice not provoke more of a public outcry from defenders of liberty than it now does (at least in the UK)? One depressingly plausible answer is that it is tempting to assume that those remanded in custody are (probably) guilty as charged; their status is not so much that of "unconvicted"

DOI: 10.4324/9781003276029-6

as that of "not yet convicted"; the presumption of innocence is tacitly assumed to have been defeated by their arrest and charge.

The formal grounds for such detention have nothing to do with punishment (the person has not been convicted),[7] and everything to do with preventing risk. Under English law, for instance, a defendant should be detained if the court is satisfied that there are "substantial grounds for believing" that, if released, he would "fail to surrender to custody"; or "commit an offence while on bail"; or "interfere with witnesses or otherwise obstruct the course of justice"; or "be likely to cause physical or mental injury [or the fear of it] to an associated person".[8] The main grounds for detention thus concern the risk that the defendant would, if released, commit an offense – either connected to his trial, such as absconding, or interfering with the trial process, or any offense at all. How could such detention be justified? If it can be justified, what kinds of consideration should courts take into account in assessing such risks?

To focus on the question of principle, we should not take it that what is to be justified is pre-trial detention as currently practiced in, for example, England or the US. Such destructive detention, imposed on so many people, cannot be justified, but we can imagine improvements that would at least significantly mitigate that destruction.[9] We could tighten the criteria for detention, and improve the process by which decisions on detention are made; we could ensure proper legal representation and advice for defendants, to resist courts' tendency to accede too quickly to prosecutors' requests for detention; we could put more resources into alternatives to custody – although if they constrain defendants, they raise issues of principle akin to those raised by detention; we could improve conditions of detention and ensure that those who are detained have access to useful activities; we could improve connections to the outside world, including families and lawyers; we could shorten the time spent in detention and mitigate its effects on housing and employment. Such improvements would be expensive and politically unpopular, but they would weaken the more contingent objections to pre-trial detention, allowing us to focus on the principled objections to any such detention, however "civilized" it is made.

One familiar, simple answer is that pre-trial detention is in principle unjustified: the court must presume the defendant to be innocent of the crime charged; and a core liberal principle dictates that we must not detain innocent people to prevent crimes they might commit if left free. A liberal state should respect the freedom, the autonomy, of its responsible citizens; it should therefore not detain them (which denies their freedom) merely on the ground that they might exercise that autonomy to wrong others. It can warn them of the consequences of doing so and intervene to prevent them carrying through a criminal enterprise on which they have embarked, but it must not seek to pre-empt the exercise of their autonomy. However, such an answer is too quick: for we do sometimes detain innocent people for preventive reasons and think that we are justified in doing so without compromising liberal principles;[10] and once we enter the criminal process, as suspects or defendants, the presumption of innocence must be to some degree qualified, or else there could be no criminal process.

As to the first point, we should note three common practices: the detention of those who are mentally disordered, to prevent harms that they might otherwise cause to themselves or to others;[11] the detention of suspected terrorists;[12] and the compulsory quarantine of those who might be infected with a dangerous disease. We should also note the ways in which those who have been convicted of criminal offenses might be subjected to detention or other kinds of restriction that look preventive rather than punitive – for instance, to detention beyond the term that is required as punishment on the grounds that they present a continuing danger to others.[13] These all, however, differ in significant ways from the pre-trial detention of criminal defendants. First, crucial to the justification of detaining the mentally disordered is that they lack the rational capacities necessary for responsible agency, but the liberty that liberals value (and that pre-trial detention infringes) is the liberty to live an autonomous, responsible, life. Second, the detention of suspected terrorists is highly controversial, but can be most plausibly justified (if it is justifiable at all) as an emergency measure in a context normatively akin to war, since we could see terrorists as engaged in a war against the polity; although the wartime detention of suspected enemies is itself controversial, it is a different controversy from that concerning the routine peacetime pre-trial detention of defendants. Third, those subjected to compulsory quarantine are detained, if they are detained at all,[14] because they might directly endanger others as soon as they come into contact with them, whereas those detained pre-trial are dangerous only in virtue of crimes they might voluntarily commit if left free; their detention is likely to be so short that there is no danger of serious impact on their lives; and it does not reflect suspicion of criminal proclivities.[15] As for those convicted of criminal offenses, there is much to be said about the (un)justifiability of preventive detention that lasts beyond what is justifiable as deserved punishment,[16] but my concern here is with the preventive detention of those who have been charged with but not yet tried for an alleged crime: the question is not whether a criminal conviction makes a normative difference that could legitimize preventive detention, but whether being charged with a crime can make such a difference.[17]

As to the second point, about the presumption of innocence, it plays its most familiar role within the criminal trial: the court that tries the defendant must begin with no assumption that he might be guilty but must treat him as innocent of the charge until the prosecution proves his guilt.[18]

There is controversy about whether we can usefully talk of the, or a, presumption of innocence outside the confines of the trial – as applying to the wider criminal process, or to the state's dealings with its citizens:[19] but the key point here is that it cannot apply in its strict form to police and prosecutorial activities in investigating crimes and bringing charges. The police must be able to treat someone as a suspect – someone whose innocence has been brought into doubt; otherwise, they could have no good reason to investigate or question him. Prosecutors cannot be expected to bring charges only if they are satisfied that the suspect's guilt has been proved (let alone proved beyond reasonable doubt): such proof is a matter for the court and will depend on what emerges during the trial. The most that can be demanded is something like the English "evidential" test; is there "sufficient evidence to provide a realistic prospect of conviction?"[20] If they had to presume everyone who came to their attention to be innocent, they could never charge anyone: how could it be right to charge with a crime someone whom I presume to be innocent of that crime? In any functional system of criminal law, there must be room for a distinctive normative role of suspect: even if we are generally to be presumed innocent,[21] it must be legitimate for the police to suspect us of committing a crime, given evidence to make such suspicion reasonable, and to investigate and question us; and for a prosecutor to charge us and bring us to trial. In acquiring that role, we acquire new responsibilities and liabilities: even if we have no duty to assist the police or play an active role in our trial,[22] we are liable to be arrested and questioned by the police, we are required to appear for our trial; our normative position changes, because the presumption of innocence is qualified.

We therefore cannot simply assert that pre-trial detention is unjustifiable because it is not consistent with the presumption of innocence: we must ask more carefully what difference(s) being charged with a criminal offense can make to our normative position – and whether one of those differences is that we can justifiably be detained pending our trial, given our status as suspect and defendant. In the remainder of this paper, I will take for granted (without trying to further explain or justify) the liberal principle that generally forbids the preventive or pre-emptive detention of responsible agents, and ask whether it can be qualified in its application to those awaiting trial as criminal defendants. In Section 4.2, I will reject three suggestions about how the fact of being charged can bear on the justifiability of pre-trial detention. In Section 4.3, I will offer a different, more plausible suggestion, based on the distinctive responsibilities that define the role of criminal defendants: these can, I will argue, justify imposing special constraints, even including preventive detention, on those who are awaiting trial. Finally, in Section 4.4, I will discuss the kinds of evidence that can properly ground a detention-justifying prediction of risk.

4.2 The Normative Significance of Being Charged?

We have noted the familiar liberal slogan that a state must not detain a responsible citizen simply on the grounds that he is judged to be criminally dangerous – likely to commit even a serious crime. Does the fact that someone has been charged with, and faces trial for, a crime make a normative difference: can it defeat or qualify that slogan? One answer is that the fact of being charged has no normative significance in this context: if we are justified in detaining a defendant on the basis of a prediction that there is an N% chance that he will commit a crime (or a serious crime) if left free, we would also be justified in detaining someone who is not a defendant of whom a similar risk assessment is true. We might then argue, as I once argued (Duff 2013), that such pre-trial detention is therefore unjustified in the same way as any other kind of preventive or pre-emptive detention of responsible agents; or, as Mayson, for instance, argues (2018, 2022), that we should be ready to detain both defendants and non-defendants for preventive reasons. However, my interest here is in attempts to show that that fact does have normative significance for what we may demand of or impose on defendants – and that that difference might ground a justification for pre-trial detention.

First, we might note that that fact does make a contingent difference in predictions of future crime. For in a properly functioning criminal justice system, most who are charged are in fact guilty: if someone has been charged, it is therefore likely that he committed the crime; and since past criminal conduct is a predictor of future criminal conduct, the fact of having been charged increases the likelihood that the person will offend in future. But that is not to say that a charged defendant is more likely to offend than anyone not currently facing charges: if the prediction is based only on (probable) past criminal conduct, we have the same reason to suspect that anyone with a prior criminal conviction will commit further crimes.

That is true, at least, if we think only about crimes in general, or of the same type as that with which this defendant is charged; it is not true if we focus on the other typical grounds for pre-trial detention – the (perceived) risk that the defendant will fail to appear for trial or will interfere with the course of justice (for instance by threatening or bribing witnesses). For it is the fact of facing trial that makes such crimes possible and gives the defendant reason to (be tempted to) commit them. Indeed, it can give such a reason and create such a temptation, even for a defendant who knows she is innocent: trials are burdensome affairs and can result in the conviction even of an innocent person. Furthermore, the motivation for such offenses obtains only for the defendant (and those who care for her, or whom she might employ) and lasts only until the trial (though there might be a motive to commit revenge or threat-fulfilling attacks on unfriendly witnesses after the trial): the court therefore has reason to ask, of any defendant, whether some preventive restrictions could be justified; and any such restrictions need last only until the trial. By contrast, if the issue is simply the commission of offenses in the future, it is not clear why defendants should be particularly liable, or why any preventive restrictions should last only until the trial. If the defendant is convicted and imprisoned, he will be prevented from committing many kinds of offense against people outside the prison,[23] whilst if he is acquitted that will preclude basing a prediction of future crime on his alleged commission of the offense charged, but there are other grounds on which we can predict future offending, by those who are but also by those who are not currently defendants, and detention based on such predictions should presumably be indefinite, until the predictions are revised, rather than ending with the trial. Thus, if the concern is with future offending in general, it is hard to see why the fact of being charged is normatively significant; it is just one kind of evidence among others.

I will discuss detention aimed at preventing trial-related offenses (absconding, interfering with the course of justice) in the following section but will focus here on detention to prevent the possible commission of other kinds of offense. A second possible reason for allowing the preventive detention of defendants awaiting trial, without seeking also to detain criminally dangerous agents in general, can be seen if we think about what such a more general practice of preventive detention would involve.

An important contingent fact about defendants is that they are available – available not just for trial, but for various kinds of assessment, including assessments of dangerousness.[24] If the state was to try to identify and detain criminally dangerous citizens in general, it would need an institutional mechanism for doing so; if we think about what that mechanism might be, we will see that it would have to involve various very disturbing kinds of intrusive official investigation, monitoring, and assessment of ordinary citizens. Focusing official assessments of criminal dangerousness on defendants can then be seen as a kind of "occasionalism":[25] the state does not try to seek out dangerous individuals in the general population, but if someone comes within the reach of the law for reasons related to criminal dangerousness (that he is charged with an offense), the state can take advantage of this "occasion" to assess whether he is dangerous.

It is hard to assess this argument. Are the criminal courts, whose primary task is to decide whether a person committed a specified offense, well equipped to assess that person's future-oriented dangerousness? Should we not instead create special tribunals whose task would be to determine dangerousness? Are there other (not unreasonably intrusive or oppressive) ways in which the potentially dangerous could come to the state's attention? But this argument can point us toward a third rationale for taking the fact of a criminal charge as significant – as opening the door to assessments of dangerousness, and detentions based on that assessment, to which citizens are not normally liable: that what is significant about the criminal charge is not that it reveals something important about the defendant's potential dangerousness, but that it makes a difference to the state's responsibilities.

Consider, for instance, Laudan and Allen's (2010) argument for a system of preventive pre-trial detention. They recommend a practice that detains, pending trial, "serial offenders (persons with more than one felony conviction within the last three years)" (2010, 34). They suggest that in the US, given plausible empirical estimates, this would result in an additional 5,671 person years of detention suffered by innocent defendants but would also prevent the commission of at least 87,000 violent crimes, which looks like a reasonable trade-off. So far, this displays a familiar style of consequentialist reasoning, which invites the equally familiar charge that the state is not entitled to "use" defendants in this way as means to the prevention of crime. But Laudan and Allen offer a further argument. The state has a responsibility to protect citizens against various evils, including crime. We must therefore weigh defendants' right to bail against the right of "innocent citizens in the community … to be protected from criminal victimization."

> Given that, if the state – having in its custody someone it believes committed a crime and who is known to have a history of criminal proclivity – nonetheless releases an individual into the community while he awaits trial, then the state bears a direct responsibility for such harm as that individual wreaks.
>
> (2010, 39)[26]

When the court releases a defendant on bail, it (and therefore the state whose agent it is) is not merely omitting to take steps to identify dangerous members of the population at large: it has been put on notice that *this* person might be dangerous; thus, the question is not whether to go out and arrest someone, but whether to release someone already in custody.

Insofar as this argument depends on an appeal to the act-omission distinction, it is not persuasive. Firstly, it is controversial whether or how it applies to the state's activities in discharging its positive responsibilities of care for its citizens.[27] Second, it is anyway unclear whether releasing a person on bail counts as an active intervention in the world to set him free (an "act") or as a refusal to continue his detention (an "omission"). But yet, there does seem (or feel) to be some intuitive force to this complaint: "You released him, even though you knew (or should have known) that he might offend." We notice similar responses when someone released from prison on parole then commits a heinous crime: the parole board should, critics complain, have been more careful. However, an analogy with parole (anyway a controversial practice) is unhelpful. A parole board is deciding whether someone serving a sentence for a crime should be released early, before the formal end of his sentence; it should release him only if it is persuaded that his release would not pose a non-trivial "risk of serious harm to the public" (Rodin 2019);[28] the presumption is that the person should remain in prison for the full term of his not undeserved sentence, unless there is persuasive evidence that he can be safely released. By contrast, in deciding whether to remand a defendant in custody, the court must presume that he should be released unless there is persuasive evidence that it would be dangerous to do so; nor can his continued detention be said to be not undeserved.

What then can we make of the thought that if a court releases a defendant who goes on to commit a (serious, violent) offense whilst on bail, it (or the state) has failed in its protective, crime-preventive, responsibilities? Not enough, I think, to justify pre-trial detention. The state has a responsibility to protect citizens from various kinds of evil, including crimes, but we must add the qualification "by legitimate means" and ask whether detaining someone who is thought to be "dangerous" is a legitimate means.[29] The state has particular responsibilities in relation to those who are within its direct control, as criminal defendants certainly are, but we must still ask what powers over those people it should have – do they include the power to detain them on the grounds that they might offend if left free? A person who is convicted of an imprisonable crime has made himself liable to imprisonment and can be detained pending his sentencing,[30] but that cannot justify detaining someone who has not been convicted or show such detention to be a "legitimate means" of preventing crime, and I have already noted the familiar liberal objection to such preventive detention: that it denies the detainee the basic liberty to which all responsible agents are entitled unless and until they forfeit it by committing an offense.[31] A defendant might well have committed the offense charged and thus forfeited that entitlement, but that has not yet been proved.

Another way to put the point is to note that in detaining a person on the grounds that they might commit crimes if left free, we are saying to them, in a drastically coercive way, "We do not trust you." We deny them that trust, that presumption of future innocence, to which citizens are entitled: to which the detainee can reply "I have done nothing to warrant such loss of trust." That reply is certainly available to one who is innocent of the crime charged, but it should also be available to the guilty person whose guilt has not yet been proved: for what warrants the removal of trust is not the very fact of offending, but the knowledge, or justified belief, that the person is guilty, and such warrant is lacking before the trial. So, we still face the question of whether and how the fact of being charged with a crime can so qualify the defendant's entitlement to be trusted – to be presumed innocent – that we can justifiably treat him as untrustworthy.

There surely is something significant in the fact of being charged with a criminal offense: significant not just for the state's responsibilities, but for the responsibilities and liabilities of the accused person; it bears on what can be demanded of her or imposed on her as she awaits her trial. I will discuss this in the following section, before discussing what kinds of evidence could ground the judgment of risk that would warrant pre-trial detention. First, however, we should note two pointers toward an answer to the question of principle.

The first pointer is found in the common idea that there is something especially heinous about crimes committed whilst on bail.[32] One explanation of that idea is that in releasing the defendant on bail, the court puts him on trust not to offend: though there was reason to suspect that he might offend, the court was ready to trust him (perhaps given the assurances provided by specified conditions of bail); in committing the crime, the defendant betrayed that trust.[33]

The second pointer lies in the suggestion that we should ask not just what we, or a court, can justifiably impose on a defendant, but what can be justifiably demanded of her: not just what can be done to her, but what she can be required to do. This suggestion reflects a wider concern to emphasize citizens' agency in relation to criminal law – to see the law not just as something imposed on us by a sovereign in relation to whom we are merely subjects, but as an enterprise that is ours, in which we have an active part to play.[34] We can see both why pre-trial detention is problematic and how it might nonetheless be justifiable, if we focus on the active duties and responsibilities that we acquire in becoming defendants in a criminal court.

4.3 Defendants and the Duty to Assure

I have argued previously (Duff 2013) that in becoming a defendant in a criminal trial, I acquire a new role, which brings with it a new set of responsibilities and burdens (as well as rights). I am now not simply a citizen who must be presumed innocent both of past crimes (except for any of which I have been convicted and for which I have paid my penal debt), and of future-directed criminal intentions:[35] I am a citizen about whose presumed innocence there is now a well-grounded doubt; for (in any decent legal system) I will have been charged and summoned to trial only if there is good evidence of my guilt. As a citizen, living under a law that is my law, I incur responsibilities to play my part in the criminal process: a legal duty to appear for my trial, and not to hinder the course of justice;

and a civic, if not a legal, duty to play my part in the trial by answering to the charge. These responsibilities are burdensome, but they are burdens I should be willing to bear, even if I am innocent of the crime charged: I must be ready to answer to my fellow citizens for my alleged criminal conduct, and thus to assist the law's enterprise of calling criminal wrongdoers to public account.

However, such burdens also create temptations. Those accused of crimes, facing trial and possible conviction and punishment, have an incentive to abscond, or to try to interfere with the criminal process: this is true both of those who are guilty, and of the innocent, if they fear the risk of mistaken conviction, or the further burdens of a trial. Their fellow citizens might therefore reasonably fear that they might try to abscond, or to pervert the course of justice; that fear will reasonably be greater the more serious the charge that the defendant faces. We can therefore now require defendants to reassure us that they will appear for trial and will not interfere with the criminal process: it is a familiar feature of our social lives, especially when we are dealing with relative strangers, that we might need to (re)assure each other of our bona fides. Such reassurance might initially be merely verbal: I promise the court that I will appear for my trial, and will not interfere with witnesses.[36] But words are notoriously cheap; at least when the alleged crime is serious, the stakes accordingly high, and the temptation to abscond or interfere accordingly stronger, it might be reasonable to require something more.

The most familiar "more" that we might require is monetary bail: the defendant must put up a sum of money that will be forfeited if she does not appear for trial (or a friend might put up the money for her). This provides an obvious disincentive to flight and is an obvious way in which defendants can provide further assurance: I put my (or my friends') money where my mouth is. The equally obvious objection to monetary bail is that it discriminates against the poor and indigent, who cannot raise the necessary funds; many of those who are in prison awaiting trial are detained simply because they cannot afford bail.[37] Could we operate a more equitable bail system that calibrated bail to the defendants' means? Perhaps not: those who are most indigent might well not be able to raise even a very modest amount. But this is one among several issues that I cannot pursue here; I will instead look briefly at some other kinds of requirements that might be rationalized as matters of assurance.

Defendants might be required, as a condition of bail, to report to the police regularly, or to surrender their passports, or to stay away from particular people or locations where they might seek to interfere with the course of justice.[38] By imposing such conditions, the court seeks to make it less likely that the defendant will succumb, or be able to succumb, to the temptation to flee or to interfere; by accepting such conditions, defendants assure the court and their fellow citizens of their readiness to play their proper part in the criminal process. I cannot discuss the range of bail conditions that can be justified in this way, but we should note some considerations that bear on their justifiability. We must ask how constrictive or intrusive they are – how far they impinge on the defendant's normal life; how important it is to secure such assurances (which depends in part on the seriousness of the charge); whether other, less restrictive conditions would be as efficient.[39] They will be most easily (which is not to say easily) justified if they do not seriously constrict the defendant's life and activities, and if they are at least relatively indiscriminate – if they do not say to the particular defendant "We do not trust *you*, in particular, to appear for trial and not to interfere with the course of justice."[40]

Suppose we can get this far and justify a system of pre-trial requirements and constraints that apply to all defendants – or to all those facing charges of a certain seriousness; that for a limited period impinge on their freedom, but leave them largely able to continue with their ordinary lives; and that can be justified to them as proportionate burdens which they should accept (whether guilty or innocent) as being necessary to reassure their fellow citizens that they will respect the criminal process. We now face three further questions –

- Can we in the same way justify pre-trial detention for certain types of defendant?
- Can we be justified in making detention, or other constraints, more selective?[41]
- Can such constraints be justified not only in the way we have discussed so far, as means of assuring the defendant's attendance at trial and non-interference with the process, but also as means to prevent his commission of offenses unrelated to the trial?

Though I will focus on detention, we must bear in mind that requirements that do not involve physical detention behind prison walls can still be just about as constricting as being locked up (see Noorda 2015).

The two obvious problems with detention are, first, that even if its conditions are vastly improved, it still radically separates detainees from their normal lives. Second, even if we can mitigate the first problem by making the walls of the place of detention more porous, the fact remains that in locking someone up, we display a more radical lack of trust in them, and thus a more radical infringement of their responsible agency. Other pre-trial requirements short of detention say to the defendant "We trust you to behave ['behave' as a shorthand for 'appear for trial and not try to interfere with the process'] so long as you accept and undertake these precautionary provisions"; pre-trial detention says "We do not trust you to behave," which is a serious insult – at least to defendants who intended to behave anyway. So our first question is whether there are kinds of crime that, given their seriousness, could warrant such mistrust: could we say, for instance, that the temptation to abscond or to interfere when charged with murder is likely to be so strong that the only adequate assurance will be detention? But even under our present law, those charged with murder can be given bail;[42] we must remember that although a murder defendant's innocence has been put into doubt, he has not been convicted, proved guilty, of the crime; nor, therefore, can we say with the requisite certainty that he has done anything to give us reason to mistrust him (for the evidence that justified charging him might not have included any suspicion-arousing conduct on his part). It is of course true that allowing those who have been charged with serious crimes to remain free pending their trial involves risk – a risk that they will fail to "behave": but that is just the kind of risk that we think we must accept as a necessary feature of a polity that treats its members as responsible agents.[43]

However, my remarks in the previous paragraph were disingenuous, since they assumed an indiscriminate, non-selective practice of detaining all those who had been charged with a serious crime. Actual practices of pre-trial detention are selective: they detain only those who present a high risk of flight or interference. Even if we cannot, for the reasons noted above, justify the pre-trial detention of *all* those charged with sufficiently serious crimes, perhaps we can justify the selective detention of "high risk" defendants; perhaps, indeed, we could argue that it would be irresponsible, a betrayal of the state's responsibility to protect its citizens and its criminal process, to fail to do so. To see whether any such practice could be justified, we must look at the criteria for detention: what could legitimately ground a detention-justifying assessment of risk? A key distinction is, I will argue, that between "There is a risk that *A* …" and "*A* presents a risk."

4.4 Criteria for Detention

Two kinds of factors are typically taken to bear on decisions about detention. In English law,[44] the court is to attend, first, to relevant facts about the defendant's prior conduct: most obviously, to his "record as respects the fulfilment of his obligations under previous grants of bail"; it is also to attend, second, to other features of the defendant's circumstances: to his "character, antecedents, associations and community ties": his "character" might be taken to consist primarily, if not exclusively, in criminal record (see Redmayne 2015); "associations and community ties" can include such matters as whether he has a job, a home, and a stable family life.

Both kinds of factor are obviously empirically relevant to assessments of the risk that a defendant will not "behave": the presence of either kind increases the probability that he will not behave. On one view, the two kinds of factor are relevant in just the same way, as bearing on the likelihood that the defendant will not behave: we are looking for an empirically well-grounded prediction of risk; we should therefore attend to all and only those factors that make it empirically more likely that he will not behave, giving each of them a weight proportionate to the degree to which it makes that more likely. This suggests an algorithmic approach: an attempt to find sound actuarial, or statistical, bases for risk assessments. Now if the purpose of pre-trial detention is to prevent the commission of offenses unrelated to the defendant's impending trial,[45] we will find no normative magic in the fact that the person being assessed is facing a criminal charge; that fact will be contingently relevant if and only if it is correlated with a higher incidence of a relevant kind of offending. However, if the court's concern is to prevent trial-related offending (absconding, interfering), the fact of being charged is clearly crucial: only a defendant can fail to appear for trial, and although others can interfere with the process on a defendant's behalf, it is the defendant who has the strongest motive to do so. But the question now is whether we should see both the kinds of factor noted above as relevant, and as relevant in the same way.

Here is a significant difference between them. We are asking whether this person can, or should, be trusted to behave. When the first kind of factor obtains, he himself has by his own prior misconduct given us reason not to trust him now: he was trusted before and betrayed that trust by misbehaving; so why should we trust him now? With the second kind of factor, we cannot say this: the fact that he is homeless, single, or unemployed may give us empirical reason to think it more likely that he will misbehave than it would be absent that factor, but it is not his misconduct, or his betrayal of trust, that gives us reason to doubt him. Whichever factor we appeal to, as the basis for a risk assessment that is to justify detaining him, we are refusing to trust him to exercise his responsible agency appropriately – we are denying him the chance to do so: but in the former case, our mistrust is grounded in his own, presumably responsible, prior failure to exercise that agency properly; in the latter case, this is not so.[46] In the former case, but not the latter, we can say that he has shown himself to be untrustworthy. In both cases, we can say that there is a risk that he will not behave, but only in the former case can we say that he presents a risk, since only then is the risk grounded in his own wrongful conduct.[47] If the latter kind of factor is highly correlated with failures to appear for trial, the answer should not be to detain the defendant, but to offer him help. If we are then asked why we should offer help (with housing, employment, or financial support) that is not available to others who are not facing criminal charges, one answer is that these are kinds of help that the polity ought to offer all its citizens; the other is that since we demand that he appear for trial, we have a responsibility to enable him to satisfy that demand.

This is not to say that any defendant who previously misbehaved when free on bail must now be detained: we must ask how serious (and recent) that past misconduct was, as bearing on how seriously it undermines present trust; how important it is that the defendant be tried in person (which will depend in part on the seriousness of the offense charged); what damage he might do to the criminal process, in particular to people involved in his case as witnesses. It is also true, of course, that if the fact of relevant prior misconduct can make a difference to what the court can now demand of a defendant, it could justify imposing special requirements short of detention – requirements more restrictive than those generally imposed on defendants.

We must be clear about the logic of the argument here. Citizens are generally entitled to a kind of civic trust: a presumption of future as well as past innocence.[48] The commission of an offense threatens that trust, but by undertaking or undergoing punishment, paying one's penal debt, that trust is taken to be restored. However, there are contexts in which a special kind or degree of trust is required, given the risks or pressures involved; one of those is the period in which a defendant is awaiting trial. The court should still usually be ready to trust defendants, if necessary subject to the kinds of requirement noted above: but prior misconduct in relation to previous trials gives the court reason to withhold that trust or to grant it only given certain special precautionary conditions. By way of partial analogy, consider driving, an activity that involves distinctive risks of serious harm. We are permitted to drive, we are trusted to drive safely, if we fulfill various legal conditions (obtaining a license, respecting traffic laws …), but we can forfeit that trust by serious (and persistent) misconduct in driving and can then lose our license to drive. Analogously, a defendant can forfeit the (conditional) trust that is normally granted to defendants by his prior betrayal of that trust – at least or especially if that betrayal was recent or persistent.

We can see the logic of this argument more clearly by turning to the other familiar bases for pre-trial detention: the risk that the defendant will commit non-trial-related offenses if he is freed on bail. One question is whether this kind of risk could ever justify detention; another is what kind or degree of risk should suffice, if we could justify such detention in principle. Some American risk assessment instruments count a person as "high risk" if there is an 8%, 10%, or 16% risk that defendants with the relevant characteristics will be arrested for a (violent) crime within six months.[49] If that degree of risk is to justify detention, it implies that we should detain about nine "innocents," who would not commit a (violent) crime if left free, to prevent one person committing such a crime – a striking inversion of Blackstone's dictum that it is better that ten guilty people go free than that one innocent person is convicted and punished (Blackstone 1753, Bk IV, ch. 27). But the question that concerns us now is about the appropriate grounds for any such risk assessment.

A first question concerns the nature of the risk to be assessed. It has to do with "danger," or "dangerousness," but just as there are, as we saw, two kinds of judgment of risk that we can make, so there are two different kinds of judgment of "danger" or "dangerousness" that might be made: that "*A* is dangerous because he might commit a violent crime if left free"; or that "There is a danger that *A* will commit a violent crime if left free." We can illuminate this difference by noticing that the second kind of judgment could also be expressed by saying "There is a danger that *A* is dangerous."

To explain. Suppose I find an object that looks very like an unexploded bomb. I might at first say "That's dangerous," meaning that it is liable to explode – that it would explode under some specifiable circumstance (it's being moved or kicked, for instance). But suppose I find that it is actually a theatrical prop: I must withdraw my claim that it is dangerous, because it clearly is not, but I can still say that there was a danger, a risk, that it would explode – that it was an actual bomb and thus dangerous.

Analogously, suppose we find a range of circumstantial factors that are correlated with a higher incidence of violent crime: these might include, for instance, gender, age, employment status, and domestic situation. We can then say that if those factors apply to *A*, there is a risk, or a danger, that he will commit a crime of violence: it is more likely that an unemployed single man of 21 will commit a crime of violence than that an elderly widow will. But this does not justify a judgment that *A* is dangerous. To say that someone is dangerous is to say that he has some disposition such that he would probably engage in the relevant conduct (in this case, commit a violent crime) in certain circumstances: but *A*'s age or employment status is not a disposition of that kind; an unemployed man of 21 might be of a peaceable, non-violent nature such that he presents no danger to others at all.[50] This point is significant because what seems intuitively plausible is that we might have reason to detain "dangerous defendants" (the title of Mayson 2018), but purely algorithmic risk assessments might show only that there is a danger that this defendant is dangerous – a less plausible ground for detaining him.

The conclusion of the previous paragraph was of course disingenuous, because central to standard risk assessments is the defendant's prior record: an important factor in justifying the conclusion that he is "dangerous" is that he has himself committed violent crimes in the past or more specifically has committed such crimes while on bail; that factor is more significant to the extent that he has thus offended more often or recently. But why is this an important factor? One answer, from advocates of algorithmic assessments, will be that it is statistically important, since it does more than other factors to increase the likelihood that this defendant will offend. But the more relevant answer is that it is important – indeed, that it provides the only legitimate kind of basis for detention-justifying assessments of risk – because it bears on whether the defendant is criminally dangerous: if we can see his prior crimes as manifesting a disposition to criminality, we can say not just that they show him to be dangerous, but that he has shown himself to be dangerous. There would of course be further empirical questions to be asked about the predictive strength of this kind of factor, but I am more interested here in its normative significance.

The mere fact of a criminal record cannot help justify pre-trial detention, for the reasons given earlier: it would be relevant if we were contemplating a practice of detaining dangerous *potential offenders*, but it does not bear particularly on those awaiting trial. What might bear on defendants awaiting trial is the fact that they had previously committed offenses whilst on bail: but this could be relevant, in line with the argument offered above about the risk that the defendant will abscond or interfere, only if release on bail requires a distinctive, enhanced, kind of trust – a trust that is betrayed, showing the defendant to be untrustworthy, not only by his absconding or interfering, but also by his commission of offenses not related to the trial. We trust you, the court must be taken to say to the bailed defendant, to behave; and "behave" is now to be taken to include not just conduct in relation to my impending trial, but a broader notion of "good behavior" (i.e., of refraining from crime).

We thus come back to the intuition that there is something distinctively heinous about a crime committed whilst on bail – perhaps because it is taken to display an especially flagrant disregard or contempt for the criminal law (though much would then depend on the particular circumstances and character of the crime): you were accused of a (serious) crime, and your response was to go out and commit another crime. I am honestly not sure whether this is an appropriate perspective, but if it is, it could help to justify pre-trial detention for those who are accused of appropriately serious crimes who have committed such crimes (or any serious crimes?) whilst on bail in the past.

4.5 Conclusion

I have argued that whilst there must, given liberal principles about respecting citizens as responsible agents, be a strong presumption against detention for those charged with criminal offenses, we can find an in principle justification for a limited practice of preventive or pre-emptive pre-trial detention. The purpose of such detention is to guard against the risk that the defendant will commit offenses (either offenses related to her trial, or any serious offense) if left free while awaiting trial, but it cannot be justified merely by an algorithmic assessment of risk. Its justification must rather start with the normative difference that being charged makes to a defendant's position, in particular with the way in which defendants acquire a particular responsibility to assure their fellow citizens that they can be trusted in this risk-laden context not to abscond or to try to interfere with the criminal process (and, perhaps, not to commit other kinds of crime whilst awaiting trial). Normally, such assurance should be accepted, they should be trusted, either without further conditions ("on their own recognizance"), or subject to certain conditions that limit but do not radically interfere with the defendant's freedom: we trust defendants to "behave" conditionally on their fulfilling those conditions. However, if a defendant has in the (recent) past shown himself to be untrustworthy, by "misbehaving" or by violating his bail conditions (or, perhaps, by committing other crimes while on bail), we have good reason to mistrust him now – to refuse to accord him even the conditional trust that we must normally accord defendants. We therefore have legitimate reason to detain him, on the grounds that he cannot be trusted (more precisely, that we cannot be expected to trust him). That reason is not by itself sufficient, but it removes the main principled objection to pre-trial detention and thus renders it in principle justifiable.

Notes

1 See https://www.gov.uk/government/statistics/offender-management-statistics-quarterly-july-to-september-2021; https://questions-statements.parliament.uk/written-questions/detail/2022-02-10/122646/

2 See https://www.sps.gov.uk/Corporate/Information/SPSPopulation.aspx (but this figure includes those who are detained pending deportation). For the US, see Sawyer and Wagner (2020).

3 See https://www.gov.scot/news/supplementary-prison-population-statistics-2019-20/.

4 https://www.fairtrials.org/articles/news/one-ten-remand-population-england-and-wales-have-been-prison-more-year/. See also Campbell, Ashworth, and Redmayne (2019, 250).

5 See Campbell, Ashworth, and Redmayne (2019, 250). The figure for non-custodial sentences might be misleading: one reason for imposing such a sentence might be that the offender has spent time in prison on remand.

6 See ECHR Articles 5, 6(2). However, Article 5(1)(c) allows for detention "for the purpose of bringing him before the competent legal authority on reasonable suspicion of having committed an offence."

7 But the likelihood of a prison sentence following conviction is relevant to an English court's decision on whether to remand in custody: see Bail Act 1976, Schedule 1, paras 1, 1A.

[8] Bail Act 1976, Schedule 1, paras 2, 2ZA, 3; on similar American provisions, see Mayson (2018).

[9] Although the mitigation would be significant only if it also addressed the ways in which current provisions for pre-trial detention have especially harsh, and discriminatory, impacts on the poor and on disadvantaged ethnic minorities.

[10] See, e.g., Mayson (2022); on prevention in relation to criminal law, Ashworth and Zedner (2014, ch. 1).

[11] For instance, in England, under the provisions of the Mental Health Act 1959.

[12] See, e.g., Cole (2009) and the provisions of s. 23 of the English Anti-Terrorism, Crime and Security Act 2001. Those provisions were finally replaced by the Terrorism Prevention and Investigation Measures Act 2011, which do not involve detention as such, but restrictive measures that do not involve physical detention can drastically constrain freedom of movement: see Noorda (2015), on "exprisonment."

[13] On English provisions, see Ashworth and Kelly (2021, chs. 9.8, 14.4). See also the case of Anders Breivik, convicted in 2012 of murdering 77 people: he received the maximum sentence of 21 years' imprisonment – which caused controversy in Norway and surprise elsewhere; but Norwegian law also provides for the continued detention, beyond their punishment, of those judged to be highly dangerous.

[14] If, that is, they are kept behind locked doors, rather than being instructed to stay at home on pain of being liable to punishment if they do not: but this distinction is far from sharp – consider, for instance, so-called quarantine hotels with no locked doors, but with security personnel patrolling.

[15] Further, in many cases (as with quarantine for those entering a country), the judgment is not that this person is dangerous (because infected), but that there is a risk that he is dangerous; and in such cases, the detention is non-discriminatory – *everyone* entering the country is quarantined. See further at n. 40 below.

[16] For an interesting suggestion, see Walen (2011).

[17] This is oversimplified: as we will see, a defendant's criminal record can bear on the decision of whether he should now be detained; those prior convictions might thus have an ongoing normative effect.

[18] Compare the wording of ECHR Article 6(2): "Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law."

[19] See e.g., Netherlands Journal of Legal Philosophy (2013); Lippke (2016).

[20] Crown Prosecution Service (2018, para. 4.6); see also para. 5 on charging when there are only "reasonable grounds to suspect that the person to be charged has committed the offence."

[21] See, e.g., Floud and Young (1981, 44), on the "right to be presumed free of harmful intentions"; Nance (1994), on the "principle of civility"; Ashworth and Zedner (2014, 130–2), on the "presumption of harmlessness."

[22] I cannot discuss such putative duties here, although I am inclined to argue that in a tolerably just polity we would have civic, though not necessarily legal, duties of these kinds.

[23] Only many kinds, because some crimes can be committed from within prison; those favoring incapacitation also often ignore the fact that crimes are committed inside prisons against fellow inmates or officers.

[24] There are plenty of kinds of dangerousness other than a propensity to commit crime; and it is misleading to talk of, for instance, a persistent shoplifter as "dangerous." Though we must be cautious about the rhetoric of dangerousness, for the sake of convenience I will still talk of "dangerous" people, meaning simply those who are likely to commit serious crimes.

[25] On "occasionalism" in penal theory, see Walker and McCabe (1973, 102–3).

[26] Compare Ashworth and Zedner (2014, 69–70), on the "assumption that the state has a responsibility to prevent offences being committed by persons who are already formally 'in the system'."

[27] For this kind of argument, see e.g., Sunstein and Vermeule (2005).

[28] And see Crime (Sentences) Act 1997, s. 28(6)(b).

[29] Compare Mayson (2018, 541–4).

[30] I am assuming throughout this discussion that imprisonment can be justified as a mode of punishment. We should certainly not simply assume that to be true: but this is not the place to discuss prison abolitionism.

[31] If the court could assume that he is guilty, he could be detained pending his formal conviction, just as, once convicted, he can be detained pending his sentence. That assumption clearly informs some attitudes to pre-trial detention: hence, the frequent talk of the risk that the defendant would commit "further offences" while on bail. Even the Crown Prosecution Service's guidance on bail issues, used to talk in these terms (see now https://www.cps.gov.uk/legal-guidance/bail on "[a]ny express or implied intention to continue to offend"). Compare Lippke (2014, 118), on whether a defendant "is 'reasonably likely' to commit further, imprisonable offenses pre-trial." But that assumption is clearly illegitimate.

[32] See Sentencing Act 2020, s. 64: the fact that an offense was committed whilst on bail is an aggravating factor.

[33] Bail is thus seen as a kind of parole in the classic sense: release on a promise of good behavior.

[34] See Duff and Marshall (2016).

[35] See at n. 21 above.

[36] Thus, a defendant might be released "on his own recognizance" pending his trial – on the basis of a promise (perhaps written) to appear for trial. Note, however, that to talk of being released in this way suggests that the default would be detention – but that is just what is at issue.

[37] For some alarming US statistics (in one study 90% of remanded felony defendants had had bail set, while in another, 40% of defendants whose bail was set at $500 or lower were detained), see Mayson (2018, 492).

[38] See Sprack (2020, ch. 7. 26–33); Campbell, Ashworth, and Redmayne (2019, 244).

[39] Whether, that is, they are strictly "proportionate" to the importance of the end (the proper functioning of the criminal process) they are to serve. On the "proportionality principle," see Barak (2012, Introduction).

[40] Analogously, preventive measures such as airport security checks, or requirements during a pandemic to stay at home, are easier to justify if they apply to everyone; they become more problematic when they are focused on members of particular groups identified not by the way in which their conduct is suspicious, but because they fit a certain profile: see further s. 4 below.

[41] The question of selectivity applies to all kinds of constraints, but for reasons of space, I'll focus on detention. We should distinguish selectivity from making equitable exceptions. Non-selective rules may bear harshly on particular defendants, and courts should be able to make exceptions in such cases.

[42] See, e.g., Bail Act 1976, Schedule 1, para. 6ZA.

[43] Note too that if we tried to guard against that risk by detaining defendants pending trial, we would expose ourselves to a heightened risk of suffering such detention.

[44] Bail Act 1976, Schedule 1, para. 9. Compare the frightening lists of factors used in American algorithmic risk assessments quoted by Mayson (2018, 512, 568)

[45] See at nn. 48–9 below; Hamilton (2021).

[46] In the latter case, it might be his "character" that supposedly gives us reason to mistrust him: but the fact that he committed crimes unrelated to "the fulfilment of his obligations under previous grants of bail" does not give us reason to believe that he will now commit such bail-related offenses.

[47] See further at nn. 49–50 below. The same is true if he issued threats against potential witnesses – another factor that can help to justify pre-trial detention: see Campbell, Ashworth, and Redmayne (2019, 241–2).

[48] See at n. 21 above.

[49] For these examples, see Mayson (2018, 494–5); contrast Campbell, Ashworth, and Redmayne (2019, 264); they suggest that a court deciding whether to detain a defendant should have "to determine whether it is more likely than not that this defendant will commit an offence likely to result in imprisonment if granted bail."

[50] Compare the difference between "it is likely that *A* will do *X*" and "*A* is likely to do *X*" (Duff and Marshall 2021): the former might be a matter simply of statistical likelihood; the latter must be based on something particular to A's dispositions.

References

Ashworth, Andrew, and Rory Kelly. 2021. *Sentencing and Criminal Justice*. 7th ed. Oxford: Hart Publishing.

Ashworth, Andrew, and Lucia Zedner. 2014. *Preventive Justice*. Oxford: Oxford University Press.

Barak, Aharon. 2012. *Proportionality: Constitutional Rights and Their Limitations*. Cambridge: Cambridge University Press.

Blackstone, Sir William. 1753. *Commentaries on the Laws of England*. Oxford: Clarendon Press.

Campbell, Liz, Andrew Ashworth, and Mike Redmayne. 2019. *The Criminal Process*. 5th ed. Oxford: Oxford University Press.

Cole, David. 2009. "Out of the Shadows: Preventive Detention, Suspected Terrorists, and War." *California Law Review* 97: 693–750.

Crown Prosecution Service. 2018. *Code for Crown Prosecutors*. https://www.cps.gov.uk/publication/code-crown-prosecutors

Duff, Antony. 2013. "Pre-Trial Detention and the Presumption of Innocence." In *Prevention and the Limits of the Criminal Law*, edited by Andrew Ashworth, Lucia Zedner and Patrick Tomlin, 115–32. Oxford: Oxford University Press.

Duff, Antony, and Sandra Marshall. 2016. "Civic Punishmentr." In *Democratic Theory and Mass Incarceration*, edited by Albert Dzur, Ian Loader and Richard Sparks, 33–59. Oxford: Oxford University Press.

Duff, Antony, and Sandra Marshall. 2021. "Character, Propensities, and the (Mis) use of Statistics in Criminal Trials." In *The Social Epistemology of Legal Trials*, edited by Zachary Hoskins and Jon Ronson, 77–91. London: Routledge.

Floud, Jean, and Warren Young. 1981. *Dangerousness and Criminal Justice*. London: Heinemann.

Hamilton, Melissa. 2021. "Evaluating Algorithmic Risk Assessment." *New Criminal Law Review* 24: 156–211.

Laudan, Larry, and Ronald Allen. 2010. "Deadly Dilemmas II: Bail and Crime." *Chicago-Kent Law Review* 85: 23–42.

Lippke, Richard. 2014. "Preventive Pre-trial Detention without Punishment." *Res Publica* 20: 111–27.

Lippke, Richard. 2016. *Taming the Presumption of* Innocence. Oxford: Oxford University Press.

Mayson, Sandra. 2018. "Dangerous Defendants." *Yale Law Journal* 127: 490–568.

Mayson, Sandra. 2022. "A Consequentialist Framework for Prevention." *Law and Philosophy* 41: 219–41. https://doi.org/10.1007/s10982-021-09427-5

Nance, Dale. 1994. "Civility and the Burden of Proof." *Harvard Journal of Law and Public Policy* 17: 647–90.

Netherlands Journal of Legal Philosophy. 2013. "Special Issue on the Presumption of Innocence." *Netherlands Journal of Legal Philosophy* 42 (3): 167–274.

Noorda, Hadassa. 2015. "Preventive Deprivations of Liberty: Asset Freezes and Travel Bans." *Criminal Law and Philosophy* 9: 521–35.

Redmayne, Mike. 2015. *Character in the Criminal Trial*. Oxford: Oxford University Press.

Rodin, Ben. 2019. *The Parole System of England and Wales* (House of Commons Briefing Paper No. 8656). https://commonslibrary.parliament.uk/research-briefings/cbp-8656/

Sawyer, Wendy, and Peter Wagner. 2020. "Mass Incarceration: The Whole Pie." https://www.prisonpolicy.org/reports/pie2020.html

Sprack, John. 2020. *A Practical Approach to Criminal Procedure*. 16th ed. Oxford: Oxford University Press.

Sunstein, Cass, and Adrian Vermeule. 2005. "Is Capital Punishment Morally Required? Acts, Omissions, and Life-Life Tradeoffs Ethics and Empirics of Capital Punishment." *Stanford Law Review* 58: 703–50.

Walen, Alec. 2011. "A Punitive Precondition for Preventive Detention: Lost Status as an Element of a Just Punishment." *San Diego Law Review* 63: 1229–72.

Walker, Nigel, and Sarah McCabe. 1973. *Crime and Insanity in England*, vol. 2. Edinburgh: Edinburgh University Press.

5 Risks of Incorrect Use of Probabilities in Court and What to Do about Them

Anne Ruth Mackor

5.1 Introduction

Over the past decennia, forensic evidence, such as DNA, fingerprints, and gunshot residue, has come to play an increasingly important role in criminal cases. As a consequence, nowadays at least three different kinds of expertise are called upon in judicial evidential decision-making. First, expertise is required with respect to the law. The selection and interpretation of the relevant legal rules and the selection and qualification of legal facts are the “core business” of judges who are trained in law. Alongside that, expertise is required with respect to the proof of those facts that judges seek to qualify as legal facts. For example, in order to decide whether a defendant committed manslaughter, the court must determine, among others, whether stabbing was the cause of death of the victim and whether it was the defendant who stabbed the victim. In doing so, courts increasingly rely on the expertise of the forensic sciences. Forensic scientists present their findings about the evidence in terms of degrees of probability, more specifically in terms of likelihood ratios. Therefore, a third type of expertise has become increasingly important in legal cases, namely expertise in statistics and Bayesian probability theory.

In most Western countries, the judiciary consists solely of jurists. Forensic scientists, statisticians, and probability theorists are not members of the court. Instead, courts can call upon these expert witnesses as advisors.[1] The introduction of these expert witnesses in courts is meant to improve judicial evidential decision-making, but it also introduces the risk of judicial misunderstandings and misapplications of forensic findings. Accordingly, we are confronted with the paradoxical situation that the introduction of expert witnesses in court can cause the quality of the court’s evidential decisions to deteriorate instead of improve.

The case of *R v. Sally Clark* is perhaps the most infamous example of the adverse effects of forensic experts using statistics and probability theory in court.[2] A medical expert witness made statistical mistakes that

DOI: 10.4324/9781003276029-7

went unnoticed by the lower court and the first court of appeal and these resulted in the wrongful conviction of Mrs. Clark. In an official statement, the Royal Statistical Society said: “The case of R v. Sally Clark is one example of a medical expert witness making a serious statistical error, one which may have had a profound effect on the outcome of the case. [...] The Society urges the Courts to ensure that statistical evidence is presented only by appropriately qualified statistical experts” (2001). The advice of the Society may be one step in the right direction, but it does not seem enough to prevent courts from making mistakes. Even if statistical evidence is only presented by qualified experts, judges and other legal factfinders still face the problem of correctly interpreting and applying these findings in their evidential decision-making.

In this chapter, I discuss the risks of the incorrect use of probabilities in court and the question of what to do about them. I examine the nature of these risks and the intricate interplay between risks and responsibilities within the rule of law. For practical reasons, I restrict myself to a discussion of criminal law and I take most of my examples from Dutch criminal law because it is the system that I am familiar with. However, my analysis is meant to be relevant for other legal systems, in particular for continental systems in which judges, not juries, are the factfinders.

In the next section, I first briefly discuss an experiment to show how the use of probability theory in court can easily result in fallacious judicial evidential reasoning. Subsequently, I analyze the nature of the risks of misinterpretation and misapplication of probabilistic findings in more detail. I also discuss the question who should be responsible for reducing those risks. Next, a large part of this chapter is then devoted to a discussion of three possible solutions to reduce the risk of probabilistic errors in court. I pay close attention to the demand that all solutions must be in accordance with the rule of law in general and with the demands of a fair trial in particular. The first and most evident solution is more judicial training in probability theory. However, this does not seem sufficient to reduce the risk of errors, or so I shall argue. The second possible solution is the introduction of probability experts as what I call “probability clerks.” This solution seems to suffice in evidentially simple cases, but not in evidentially complex cases. I argue that we need a third and more radical solution, viz. the introduction of what I call “probability judges”, at least in evidentially complex cases.

My main conclusions are that we do not know the number and the severity of probabilistic errors in court, but that we have reasons to worry about them, in particular about the risks of miscarriages of justice. My main recommendations are, first, that empirical research be done to investigate the precise number and the nature of the risks and, second, that experiments with probability clerks and probability judges be done in order to empirically test whether their introduction can reduce the number and severity of probabilistic fallacies in court.

5.2 Probabilistic Reasoning in Court: An Example

I start with an example that illustrates the role that probabilistic reasoning nowadays plays in forensic reports. It offers insight into one type of probabilistic misunderstanding that these forensic reports can cause in judicial evidential decision-making. The example is a simplified criminal case of a robbery at a cash dispenser which I take from an experiment by De Keijser and Elffers (2012, 195–8). The two main pieces of evidence are security camera images of the robber and a report of a forensic expert who compared these images with photos of the suspect.

The expert reports that he has carried out comparative research and that he has examined if the findings fit better under hypothesis 1 than under the alternative hypothesis 2. Hypothesis 1 holds that the suspect is the perpetrator (more specifically, it holds that the perpetrator of the robbery visible on a specific CCTV image is the same person as the suspect depicted in a specific photo). Hypothesis 2 holds that the suspect is not the perpetrator (more specifically, it holds that the perpetrator of the robbery visible on the CCTV image is not the same person as the suspect depicted in the photo). The expert reports that the findings based on the selected visual materials of the facial comparison are much more likely when the person depicted is the same person (hypothesis 1) than when they are different persons (hypothesis 2).

De Keijser and Elffers asked judges, defense lawyers and experts what they can correctly derive from this report. Among others they asked the participants whether the following is a correct interpretation of the conclusion of the expert: "It is much more likely that the suspect is the person on the images from the security camera than someone else is the person on those security camera images" (De Keijser and Elffers 2012, 198). More than 88% of the judges and lawyers and more than 63% of the experts believed this conclusion to be correct (De Keijser and Elffers 2012, 199–200).[3] Unfortunately, however, it is false. The expert reports on the probability that one will find the evidence, *given a particular hypothesis*, but most participants – including a majority of the forensic experts – interpret the statement as a report on the probability that a particular hypothesis is true, *given the evidence*. The mixing up of these probabilities is called the prosecutor's fallacy (Thompson and Shumann 1987).[4] An even more worrisome finding of De Keijser and Elffers is that more than half of the judges and defense lawyers and 85% of the experts claimed to have a perfect or near-perfect understanding of the forensic conclusions presented to them (2012, 201–2). In other words, not only did a majority of the participants misinterpret the report, but many of them were also blind to their own lack of understanding.

One needs to have basic knowledge of Bayesian probability theory to understand the prosecutor's fallacy. For the purposes of this chapter, however, there is no need to go into the details of probability theory. It suffices if the reader has an intuitive grasp of the nature and of the potential far-reaching consequence of this type of mistake. Let me therefore present a simple example. Compare the following two questions. First, what is the conditional probability that a randomly chosen mammal has *four legs*, if (condition) it is a *cow*? The probability that a randomly chosen cow has four legs is quite high. Second, what is the conditional probability that a randomly chosen mammal is a cow, if (condition) it has *four legs*? This probability seems very low. This example helps to understand that these two conditional probabilities can diverge dramatically. Now we see more clearly that it is one thing to say that there will probably be a match between the findings if the defendant is the perpetrator, but quite a different claim to say that the defendant is probably the perpetrator if there is a match. If courts mix up these probabilities, like the participants of the experiment of De Keijser and Elffers did, they run the risk of reaching incorrect conclusions about the probability that the defendant committed the crime and therewith they run the risk of committing a miscarriage of justice.

5.3 Responsibility and Risk

I began this chapter by referring to the risks of incorrect use of probabilities in court and the question of what to do about them. We have just seen that mistakes in the interpretation and application of probabilistic statements can result in fallacious argumentation, false conclusions, and – in the worst case – miscarriages of justice.

Risk is standardly defined as the statistical expectation value of an unwanted event that may or may not occur, or as the product of the probability that an event will take place and the degree of "unwantedness" or severity of that event (Hansson 2018). Accordingly, to say that a risk is high can mean that both the probability and the severity of an event are high, that the probability of the event is high but the severity low, or that the probability is low but the severity high. However, if the severity is deemed very low, we no longer speak in terms of risks.

5.3.1 First-Order Risk and Second-Order Risk

The experiment of De Keijser and Elffers and the other literature I referred to suggest that the probability that judges and other legal factfinders make mistakes in probabilistic reasoning is quite high. This holds in particular for the prosecutor's fallacy. Next to that, several other probabilistic fallacies have been distinguished in the literature. For example, Dahlman (2018) has distinguished ten types of probabilistic errors, such as base rate neglect, underestimating the combined strength of concurring evidence and dependence neglect.[5]

However, as far as I know, we do not know how often courts make probability mistakes, nor do we know how severe the consequences of these mistakes are. There are indications that the prosecutor's fallacy is made regularly, and the same seems to hold for the base rate neglect and for the underestimation of the combined strength of weak evidence. However, even if we assume that courts regularly make these mistakes, then we still do not know how often these fallacious inferences result in miscarriages of justice.

In other words, we are not only confronted with first-order risks, that is, with the risk of probability mistakes in court and the risk that these mistakes result in miscarriages of justice. Because we are ignorant both about the probability of probability mistakes and about the severity of their consequences, we are uncertain about the magnitude of the first-order risk. Therefore, we are also confronted with second-order risks.[6] If we incorrectly believe the first-order risk is high, we will spend too much effort on preventing mistakes. Conversely, if we underestimate the first-order risk, we end up taking insufficient preventive measures. In conclusion, we need empirical research to determine the first-order risk of judicial probability mistakes and therewith to lower the second-order risk.[7]

5.3.2 Risk versus Uncertainty; Objective versus Subjective Probability

Given that we are uncertain about the number as well as about the severity of the unwanted effects of probability mistakes, some readers might want to object to my use of the term "risk." Following Knight's distinction between risk and uncertainty, they could argue that for lack of quantifiable probabilities about probability mistakes, I should speak in terms of uncertainty. Knight states: "To preserve the distinction [...] between the measurable uncertainty and an unmeasurable one we may use the term 'risk' to designate the former and the term 'uncertainty' for the latter. [...] We can also employ the terms 'objective' and 'subjective' probability to designate the risk and uncertainty respectively, as these expressions are already in general use with a signification akin to that proposed" (Knight 1921, 233).

Let me make two brief remarks on this issue. First, the probabilities that we are interested in are as yet unknown, but they are not unknowable in principle. Empirical research could deliver the statistical information we need and, in fact, one of my recommendations is that research be done to gather that information. Second, as the quote makes clear, Knight does not only use the terms risk and uncertainty, but he also refers to the distinction between objective and subjective probability. Let me explain the difference between these two types of probability (Hacking 2001, 132–7).[8] If a coin has been tossed 100 times and landed heads 45 times, then the frequency of heads of this coin is .45. If we say that (in the long run) the probability of getting heads with this coin is .45, we seem to be using an objective or frequency-type probability. However, if we want to assess the probability that the coin will land heads the next time I toss it, talking in terms of frequency-type probability does not make sense. On a single occasion, a coin will either land head or tails. In a single case, we can only use a subjective or belief-type probability. We can say that our degree of belief that the coin will land heads is .45, even though the reason to have this degree of belief is the information I have about the frequency that the coin has landed heads in the past. The same holds for the example about the probability that the CCTV images and the photo of the defendant match if the defendant is the perpetrator. This too is a belief-type probability. Judicial decision-making is about single cases. Therefore, judicial decision-making is about belief-type, that is, Bayesian, probabilities. Therefore, in judicial decision-making, the distinction between uncertainty and risk is not a fundamental or principled distinction, because on a Bayesian account, all probabilities, even those informed by "objective" frequencies, are subjective.

5.3.3 Material Risks and Epistemic Risks

Let us return to the risk of probability mistakes in judicial decision-making. I have argued that we are uncertain about the probability and the severity of probability mistakes made by courts. However, we can be more precise about the nature of the risks involved in the judicial interpretation and application of probabilistic statements. We can distinguish two types of risk. First, there is the risk that courts make unsound inferences and that they, as a consequence of these unsound inferences, come to adhere to false beliefs. Second, these false beliefs can have further adverse consequences, incorrect decisions – in the worst case miscarriages of justice – and their executions.

When we talk about risks, we often focus on the latter type of risk, i.e., on the material or practical risk of (the execution of) wrongful convictions and wrongful acquittals that can follow from unsound inferences and false beliefs. The reason to call them material or practical risks is that they put our value of practical rationality at risk because and to the extent that they are about making and executing or implementing legally, morally, and politically wrongful decisions. However, the events of making unsound inferences and adhering to false beliefs are not only unwanted because of their undesirable practical consequences. Committing fallacies and adhering to false beliefs are also in themselves unwanted, because and to the extent that they conflict with our value of epistemic or theoretical rationality.[9] Epistemic rationality is not only a scientific value, but also a fundamental value of criminal trials since these trials aim not merely at a procedural truth, but primarily at the material truth. Accordingly, if a court, by moral luck, makes a correct legal decision that is based on flawed reasoning and/or false beliefs, epistemic injustice is nevertheless done. Therefore, the risks of committing fallacies and adhering to false beliefs are called epistemic risks.

Accordingly, we can distinguish between practical or material risks of wrongful decisions and their execution on the one hand and theoretical or epistemic risks on the other. At least three different types of epistemic risk play a role in the judicial interpretation and application of probability statements in legal cases. The first epistemic risk is the second-order risk mentioned above, i.e., the fact that we are deeply uncertain about the nature and the magnitude of the first-order risk, viz. about the probability and the severity of courts making probability mistakes. The second epistemic risk is the risk of making unsound inferences, regardless of whether these inferences result in false beliefs. The third epistemic risk is the risk of actually entertaining false beliefs as a consequence of unsound reasoning.

5.3.4 Who Should Be Responsible?

Before turning to the question whether it is possible to reduce these risks, let us briefly discuss the question of who is or should be responsible for assessing and reducing them. At first sight, it seems quite logical to say that the judiciary as a state organ is responsible for assessing the quality and the quantity of the risks of making probability mistakes in judicial evidential reasoning and that both the judiciary as an organization and individual judges are responsible for minimizing these risks. Both the judiciary as a whole and individual judge must ensure that courts are competent to perform their task of evidential decision-making. Like other professionals, they need to see to their own training and ask for advice if they lack specific expertise. Accordingly, if they lack competence in forensic sciences and probability theory, they should get more training and/or advice.

However, the judiciary and judges functioning in the rule of law differ from other professionals and professional organizations in some crucial respects. If professionals lack competence for a particular task, either they will not perform the task themselves and refer clients to another professional who is competent, or they will collaborate in a team of professionals so as to make sure the team as a whole has the required competence. Judges, however, cannot operate in a similar way. First, they are not allowed to refuse to decide a case; that would be a denial of justice.[10] Second, to ensure their independence and impartiality, judges are not allowed to collaborate with other professionals when deciding a case. Courts can ask experts for advice, but they have to make the decision on their own: experts are allowed, as advisors, in the court room but not, as decision makers, in the council chamber. For the same reason, even though courts can discuss general characteristics of a case with "outsiders" and ask for general advice, they are not allowed to discuss their envisaged decisions in specific cases.

These restrictions have ramifications for the nature of possible solutions. In the first place, the judiciary can only take measures within the confines of the law. Secondly, although the legislator can change the law, it must see to it that the solutions are in accordance with fundamental human rights, in particular with the right to a fair trial as it is laid down in constitutions and in international treaties like article 6 of the European Convention on Human Rights (ECHR) and article 14 of the International Covenant on Civil and Political Rights.

5.4 Three Possible Solutions

In this section, I discuss three possible solutions that are in accordance with the rule of law and the right to a fair trial. The first solution does not demand any adaption of (Dutch) criminal law, the second demands a slight adaptation, and the last is the most revisionary proposal. As far as I can see, they are the only feasible solutions within the bounds of the rule of law and the (Dutch) system of law.

5.4.1 Training

The first and most obvious solution to reduce the number of probability mistakes is to provide judges with more education in probability theory. Through these trainings, judges can acquire passive understanding of probability theory, in particular of Bayes' rule, and of important concepts such as the prior probability and the likelihood ratio. However, it seems much more difficult and possibly too time-consuming to acquire the ability to reason actively and correctly with probabilities and to detect errors in probability reasoning of oneself and others. Teaching judges basic understanding of probability theory is definitely necessary, but it does not seem sufficient to prevent them from making serious probabilistic mistakes.[11]

5.4.2 Probability Clerks

Another possible solution is the appointment of probability clerks, i.e., assistants with specific expertise in probability theory. This solution fits nicely with recent developments in the Dutch judiciary. In 2012–13, a Forensic Support pilot has been conducted in a number of courts. The pilot consisted of the appointment of forensic assistants, generalists with a master's degree in forensic sciences (Raad voor de rechtspraak 2014, 5). In 2014, the pilot had been evaluated positively, and it resulted in the appointment of forensic assistants at all criminal courts, both lower courts and courts of appeal (Raad voor de rechtspraak 2014).[12]

The task of the forensic assistants is, among others, to prepare the forensic parts of criminal files and to answer clarificatory questions from judges about forensic reports and about the hearings of forensic experts. By analogy, assistants who have obtained a master's degree in probability theory or statistics could be appointed as probability clerks. The advantage of appointing probability experts as assistants next to expert witnesses is that probability clerks can explicate probability arguments in the forensic reports, not only in the preparation for the hearing, but also after the hearing and in the council chamber.

Of course, we do not know whether the introduction of probability clerks would be as successful as the introduction of forensic assistants. A pilot should be conducted to find out whether probability clerks can help to reduce the number and the severity of probability mistakes made by courts. However, there are several reasons why we should doubt that their appointment suffices in evidentially complex cases. I mention three limitations in particular. First, probability clerks can prepare questions for the court, but as clerks they are not allowed to ask (follow-up) questions during trial at the hearing of the experts. At the crucial moment of the hearing, the judge has to ask the proper questions without assistance. Second, although probability clerks can, like forensic assistants, be present in the council chamber to answer clarificatory questions, they are not allowed to participate in the deliberations. Moreover, and this is a third limitation, as clerks, they are not allowed to give their own interpretations of probability statements because they are not experts in the sense of the law (Raad voor de rechtspraak 2014, 5).[13]

5.4.3 Probability Judges

The limitations on the role of probability clerks suggest that we need probability judges, at least in evidentially complex cases.[14] The introduction of probability judges might sound problematic. Readers might worry that it conflicts with the rule of law and with the continental view that jurists have a monopoly on the judiciary and that they have it for good reasons.

A third possible objection is that my plea for probability judges opens the floodgates to the introduction of many more types of expert judges. Let me discuss these three worries in turn.

5.4.3.1 The Monopoly of Legal Professionals in the Judiciary

First, it should be noted, that even in continental countries the juridical monopoly on the judiciary has never been complete. Most countries have lay judges and/or some mixed chambers of the court. Second and more importantly, we can observe an analogy between the current situation and the situation at the beginning of the 19th century when legal professionals obtained the monopoly on the judiciary. By the end of the 18th century, law had become so complex in many Western countries that it was no longer deemed sufficient that courts consisted of lay judges who let themselves be advised by a legal professional. The turning point in the Netherlands was the Dutch Code on the Judicial Organization of 1827 that ordained that all judges of lower courts, courts of appeal, and the supreme court should be jurists with specific legal training and competence (Van Boven 1990, 267–70).

The analogy between the 1820s and the 2020s can easily be seen. In those days, it was the law; in our days, it is not only the law, but also the assessment of evidence, especially of forensic evidence that has become too complex to be handled by lay persons. The interpretation and application of evidential findings has become so complex that it demands specific probabilistic competence. As in the 1820s, there are reasons to believe it is no longer sufficient that the judge is advised by experts, but that probability experts should themselves be members of the court.

5.4.3.2 The Right to a Fair Trial

This takes us to the question whether probability judges are in accordance with the role of the judiciary in the rule of law. The fundamental task of the judiciary lies not only with the correct application of material law, but also or even primarily with safeguarding a procedure that ensures a fair trial. Article 6 (1) ECHR, for example, states that "In the determination of his civil rights and obligations or of any criminal charge against him, everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal established by law." The introduction of probability judges does not seem to conflict with the demands of a fair, public, and timely hearing or with the independence and impartiality of the court. Moreover, another crucial aspect of a fair trial is that, in the end, the court delivers an understandable and properly reasoned judgment. As I have argued in foregoing sections, this is exactly what is at stake in evidentially complex cases. Stated yet differently, we need probability judges to fulfill a fundamental requirement of a fair trial, viz., that courts deliver properly reasoned judgments.

Another important aspect of a fair trial is the adversarial principle. One aspect of this principle is the requirement that parties have had sufficient opportunity to react to all the evidence and the arguments. This implies that judges cannot discuss any insights, reasoning, or information in the council chamber or present them in their final decision if they have not been discussed during the hearing. It is the task of the legally trained presiding judge to guard this and other important aspects of a fair trial and the probability judge should be trained to act in accordance with this fundamental adversarial principle.

5.4.3.3 Opening the Floodgates?

A final objection to my plea for probability judges is that it opens the floodgates to the introduction of many more types of expert judges. However, I believe that this objection fails too. First, it should be noted that the expertise of probability experts differs substantially from the expertise of other experts. For one thing, probability experts do not have, like forensic experts, "substantive" expertise about the material or "underlying" facts. Probability theorists are experts with respect to reasoning with and about probabilities. Since all forensic experts make probabilistic claims, probability judges can assess the quality of their probabilistic arguments. In this respect, probability experts are on a par with legally trained judges who also lack substantive expertise and who are "only" experts in reasoning with and about legal rules. In seeing to it that cases are decided in accordance with legal rules, legally trained judges are the guardians of practical rationality (Schauer 1993). Similarly, in seeing to it that cases are decided in accordance with the rules of probability theory, probability judges would be the custodians of theoretical rationality.

Therefore, or so I conclude, my proposal does not open the floodgates to a whole array of other types of expert judges. On the contrary, given that probability judges can strengthen both the critical evaluation of forensic reports and the evidential reasoning of the court itself, my proposal is perhaps the best way to manage and improve the ever-increasing contributions of forensic experts in the administration of justice, at least in evidentially complex cases.

5.5 Conclusions and Recommendations

In this chapter, I have discussed the risks involved in the judicial interpretation and application of probability statements. I have argued that there are material risks and three types of epistemic risks. First, for lack of solid empirical research into the matter, we are uncertain both about the probability and about the severity of probability mistakes made by criminal courts (second-order epistemic risk). However, there are several indications that the probability of judicial probability mistakes is quite high. These mistakes consist in "unsound"[15] inferences such as the prosecutor's fallacy (first-order epistemic risk) and they often result in false beliefs (another first-order epistemic risk). The ultimate risk is that they result in wrongful convictions and acquittals and their executions (first-order material risks). There are examples of wrongful convictions, such as the infamous case of R v. Sally Clark, but again we seem to lack reliable numbers.

Even though we do not know the number and the nature of judicial probability mistakes nor the number and nature of their adverse practical consequences, there are indications that judges are insufficiently competent in interpreting and applying probabilistic statements. Therefore, more training of judges is necessary. I have argued, however, that such training does not seem sufficient. Even though it is possible to teach judges basic but fairly passive knowledge of probability theory, it seems much harder and more time-consuming to teach them how to actively use probability theory themselves, to critically question experts, and to detect flaws in probabilistic reasoning.[16]

Therefore, I have argued for more far-reaching changes of our legal systems. One relatively simple change is the introduction of experts in probability theory as clerks who can explicate probabilistic statements of experts and who can help to prepare questions for experts and who can help judges to avoid making probability mistakes in their evidential argumentation. On the positive side, the introduction of probability clerks fits easily in existing legal systems. In fact, it is just one step beyond the recent successful introduction of forensic assistants in Dutch criminal courts. On the negative side, I have argued that the introduction of probability clerks might not be sufficient as a solution in evidentially complex cases.

My third and most radical proposal has been the introduction of probability judges, i.e., experts in probability theory who sit themselves as judges in mixed chambers of the court. I have argued that their introduction seems necessary in evidentially complex cases and I have argued their introduction is possible within the confines of the rule of law and without running the risk of opening the floodgates to many other types of expert judges.

Finally, I call for two types of empirical research. First, I call for research into the nature and the number of probability mistakes that are being made by criminal courts and into their material consequences. Second, I call for experiments with probability clerks and probability judges to empirically test my hypothesis that their introduction contributes to reducing the number and the severity of probability mistakes made by courts.

Notes

[1] This is sometimes called the advisory system of the judiciary. The alternative is a decision system in which judicial professionals decide cases together with professionals from other disciplines. In Western legal systems, the advisory system is the standard and the decision system the exception. One of the Dutch exceptions to the advisory system is the penitentiary chamber of the Court of Appeal Arnhem-Leeuwarden. This chamber consists of three judges and two behavioral experts (a psychiatrist and a psychologist). See De Groot and Elbers (2008).

[2] See, for example, Lagnado (2021).

[3] Perhaps the most surprising fact is that the mistakes were not only made by a majority of the judges and lawyers, but even by a majority of the forensic experts. Some forensic experts have or at least are assumed to have knowledge of Bayesian probability theory. However, it should be noted that this depends on the discipline of the experts. Not all forensic experts are trained to apply and to report in terms of Bayesian probability theory.

[4] There are some indications that the prosecutor's fallacy is frequently committed by Dutch Courts. Prakken (2018) analyzed 31 recent Dutch judicial decisions and found that the court committed the prosecutor's fallacy in 22 of them. Also see Prakken and Meester (2017). Meester and Stevens (2021) analyzed another four recent Dutch criminal cases and detected the prosecutor's fallacy, the base rate neglect, and the underestimation of the combined strength of weak evidence in them.

[5] Dahlman (2018) further distinguishes false positive neglect; wrong reference class; false dichotomy; underestimating the cumulative uncertainty in evidence chains; double-counting and double-discounting; overestimating predictive evidence. Also, see Dahlman, in preparation.

[6] See Möller, Hansson, and Peterson (2006, 422ff) about second-order risks and epistemic uncertainty.

[7] These are not the only uncertainties, however. Human beings do not only err in probabilistic reasoning, but also in many other ways and we do not know whether the risks of fallacious probabilistic reasoning are higher than the risks of other kinds of fallacies and biases. Finally, we also do not know whether debiasing measures are effective (Zenker 2021).

[8] For an application in law see, for example, Robertson and Vignaux (1993).

[9] On the distinction and the relationship between practical and theoretical rationality, see Mackor (2011) and Mackor (2013).

[10] In Dutch law, this prohibition is laid down in article 13 Wet Algemene Bepalingen [General Provisions Act].

[11] Moreover, elsewhere I have argued extensively that judges not only need to learn probability theory, but also explanation-based theories to evidential decision-making such as the scenario theory (Mackor and Van Koppen 2021; Mackor, Jellema, and Van Koppen 2021).

[12] And through personal correspondence with the Rechtspraak Servicecentrum [Service Centre of the Judiciary] July 22, 2020.

[13] The Evaluatie Pilot Forensische ondersteuning rechtbanken Straf commissioned by the Dutch Council for the Judiciary explicitly states that forensic assistants are not experts in the sense of the Code of Criminal Procedure and that they do not provide their own interpretation of the forensic evidence or the criminal case. The same would be true of probability clerks.

[14] For a more detailed analysis of the question whether and how the introduction of probability judges in the Netherlands is possible within the limits of the Dutch constitution, see Mackor and Schutgens (2022).

[15] Here, I define unsafe or unsound convictions and acquittals as decisions that are in themselves correct but based on unsound reasoning. Stated differently, in such cases, the conviction or the acquittal can be upheld, but not on the basis of the argumentation of the verdict. I distinguish these cases from wrongful convictions and acquittals in which the decision cannot be upheld.

[16] In 2022, Mackor, Dahlman, and Lagnado received a NWO research grant (number 406.21.RB.004) to develop and test a method for teaching judges to reason more rationally about evidence in criminal cases. More information at https://preventingmiscarriagesofjustice.wordpress.com/

References

Dahlman, Christian. 2018. *Beviskraft. Metod för bevisvärdering i brottmål*. Stockholm: Norstedts.

Dahlman, Christian. In preparation. A Systematic Account of Probabilistic Fallacies in Legal Fact-finding.

De Groot, Dineke, and Nieke Elbers. 2008. *Inschakeling van deskundigen in de rechtspraak: Verslag van een onderzoek naar knelpunten en verbetervoorstellen*. Research Memoranda 4(3). The Hague: Raad voor de rechtspraak. https://www.rechtspraak.nl/SiteCollectionDocuments/Inschakeling-van-deskundigen-in-de-rechtspraak.pdf

De Keijser, Jan, and Henk Elffers. 2012. "Understanding of Forensic Expert Reports by Judges, Defense Lawyers and Forensic Professionals." *Psychology, Crime & Law* 18 (2): 191–207.

Hacking, Ian. 2001. *An Introduction to Probability and Inductive Logic*. Cambridge: Cambridge University Press.

Hansson, Sven Ove. 2018. "Risk." In *The Stanford Encyclopedia of Philosophy* (*Fall 2018 ed.*), edited by Edward N. Zalta. Stanford: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/risk/

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin Company.

Lagnado, David. 2021. *Explaining the Evidence*: *How the Mind Investigates the World*. Cambridge: Cambridge University Press.

Mackor, Anne Ruth. 2011. "Explanatory Non-Normative Legal Doctrine. Taking the Distinction between Theoretical and Practical Reason Seriously." In *Methodologies of Legal Research. Which Kind of Method for What Kind of Discipline*, edited by Mark Van Hoecke, 45–70. Oxford: Hart Publishing.

Mackor, Anne Ruth. 2013. "What Can Neurosciences Say about Responsibility? Taking the Distinction between Theoretical and Practical Reason Seriously." In *Neuroscience and Legal Responsibility*, edited by Nicole A. Vincent, 53–83. Oxford: Oxford University Press.

Mackor, Anne Ruth, Hylke Jellema, and Peter J. Van Koppen. 2021. "Explanation-Based Approaches to Reasoning about Evidence and Proof in Criminal Trials." In *Law and Mind: A Survey of Law and the Cognitive Sciences*, edited by Bartosz Brozek, Jaap Hage and Nicole A. Vincent, 431–70. Cambridge: Cambridge University Press.

Mackor, Anne Ruth, and Roel Schutgens. 2022. Artikel 116 Grondwet: de inrichting en samenstelling van de rechterlijke macht. Over het monopolie van juristen en de positie van de kansendeskundige. In *Een nieuw commentaar op de Grondwet*, edited by Afshin Ellian and Bastiaan Rijpkema, 451–66. Amsterdam: Boom.

Mackor, Anne Ruth, and Peter J. Van Koppen. 2021. “The Scenario Theory about Evidence in Criminal Law.” In *Philosophical Foundations of Evidence Law*, edited by Christian Dahlman, Alex Stein and Giovanni Tuzet, 213–24. Oxford: Oxford University Press.

Meester, Ronald, and Lonneke Stevens. 2021. “Correct redeneren: wat een Bayesiaanse analyse zegt over vier recente strafrechtelijke uitspraken.” *Expertise en Recht* 14 (3): 112–21.

Möller, Niklas, Sven Ove Hansson, and Martin Peterson. 2006. “Safety Is More than the Antonym of Risk.” *Journal of Applied Philosophy* 23 (4): 419–32.

Prakken, Henry. 2018. “Kansoordelen door deskundigen: over ‘logisch’ rapporteren en wat daarbij mis kan gaan.” *Ars Aequi* 67: 740–7.

Prakken, Henry, and Ronald Meester. 2017. “Bayesiaanse analyses van complexe strafzaken door deskundigen. Betrouwbaar en zo ja: Nuttig?” *Expertise en Recht* 10 (5): 185–97.

Raad voor de rechtspraak. 2014. *Evaluatie Pilot Forensische ondersteuning rechtbanken Straf*. The Hague: Raad voor de rechtspraak. https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Raad-voor-de-rechtspraak/Nieuws/Documents/(256140063)%20EVALUATIE%20PILOT%20FORENSISCHE%20ONDERSTEUNING%20).pdf

Robertson, Bernard, and GA Vignaux. 1993. “Probability. The Logic of the Law.” *Oxford Journal of Legal Studies* 13 (4): 457–78.

Royal Statistical Society. *Royal Statistical Society Concerned by Issues Raised in Sally Clark case*, 23 October 2001. https://rss.org.uk/RSS/media/File-library/Membership/Sections/2020/Sally-Clark-RSS-statement-2001.pdf

Schauer, Frederick. 1993. *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Oxford: Clarendon Press.

Thompson, William C., and Edward L. Shumann. 1987. “Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor’s Fallacy and the Defense Attorney’s Fallacy.” *Law and Human Behavior* 2 (3): 167–87.

Van Boven, Maarten Willem. 1990. *De rechterlijke instellingen ter discussie. de geschiedenis van de wetgeving op de rechterlijke organisatie in de periode 1795-1811*. PhD thesis. Nijmegen.

Zenker, Frank. 2021. “Debiasing Legal Fact-Finders.” In *Philosophical Foundations of Evidence Law*, edited by Christian Dahlman, Alex Stein and Giovanni Tuzet, 395–409. Oxford: Oxford University Press.

Part III

Bioethical Context

6 The Failure of Luck Anti-Egalitarianism

Sven Ove Hansson

6.1 Introduction

Since the 1960s, increased knowledge about the connections between human disease and behavioral factors such as smoking, diet, and insufficient exercise has led to extensive health-promoting measures. It has also led some to put increased focus on the responsibility of individuals for their own disease. Claims have been made that certain treatments should be withheld from patients with diseases classified as self-inflicted or as lifestyle diseases. For instance, it has been proposed that for moral reasons, persons with smoking-related diseases should be excluded from insurance coverage for these diseases (Underwood and Bailey 1993) and alcoholics have lower priority than others for liver transplantation (Glannon 1998). In politics, such restrictions in access to healthcare are particularly popular among conservatives (Fierlbeck 1996; Persson 2013, 434–5). In contrast, most of the philosophers and ethicists who have promoted such restrictions describe themselves as "luck egalitarians," giving the impression that the restrictions in question contribute to a more equal society.

"Luck egalitarianism" relies heavily on Ronald Dworkin's distinction between option luck and brute luck. Option luck is "a matter of how deliberate and calculated gambles turn out – whether someone gains or loses through accepting an isolated risk he or she should have anticipated and might have declined," in other words, the outcome of the individual's own risk-taking. Brute luck is the fallout of risks that "are not in that sense deliberate gambles" (Dworkin 1981, 293).[1] According to "luck egalitarianism," society should make up for the misfortunes affecting an individual due to brute luck, but it should not compensate for disadvantages that are "in some way traceable to the individual's choices" (Cohen 1989, 914). Major proponents of "luck egalitarianism" include Eric Rakowski (1991), Richard Arneson (2000), G.A. Cohen (1989), and Shlomi Segall (2010). Major critics include Elizabeth Anderson (1999), John Harris (1995), Daniel Wikler (2002), and Phoebe Friesen (2018).

DOI: 10.4324/9781003276029-9

The purpose of this chapter is to assess some of the fundamental assumptions underlying "luck egalitarianism." Section 6.2 identifies three major such assumptions. In Sections 6.3–6.5, each of these assumptions is scrutinized. None of them is found to be tenable. General conclusions from these findings are offered in Section 6.6.

6.2 Blame Responsibility and the Argument Structure

"Luck egalitarianism" presupposes that there is a moral link between, on the one hand, the fact that a disease or injury was due to the injured person's option luck and, on the other hand, the lack of a social commitment to offering her healthcare to which she would otherwise have been entitled. This link is usually taken to be one of responsibility. We are assumed to be responsible for our own option luck (Cohen 1989, 922).

It is essential in the healthcare context to distinguish between two major types of responsibility: task responsibility and blame responsibility. By task responsibility is meant that one has to do or achieve something or get something done. By blame responsibility is meant that one morally deserves blame if something goes wrong.[2]

The two types of responsibility are often disconnected from each other in communications with patients. In many, if not most, consultations with patients, physicians, nurses, and physiotherapists offer advice or instructions for what the patient can do to promote her own health. Patients are told to take medicines, stop smoking, reduce their alcohol consumption, engage in physical activity, perform specific exercises, change their diets, etc. This means that task responsibility for these interventions is laid on the patient. However, some of these instructions are difficult to comply with, and it is not uncommon that patients fail to do so. When that happens, the patient is usually not burdened with blame responsibility. A patient who did not manage to stop smoking will not be reprehended by her physician or told that it is now her own fault if she has a heart attack. More typically, she will be offered help, for instance, nicotine patches or participation in a smoking cessation program. Thus, patients are urged to take task responsibility for actions that improve their own health, but they are not saddled with blame responsibility for their failures to follow the instructions. This may seem to be an inconsistent practice, but it is borne out by both research and practical experience showing that inciting feelings of guilt tends to hamper rather than stimulate health-promoting behavior.[3] Thus, assigning task responsibility to patients is not controversial; to the contrary, it is an essential element of good care and evidence-based medicine. The controversy concerns blame responsibility and its use as a criterion to exclude patients from treatments that they would benefit from.

Blame responsibility is a necessary requirement for "luck egalitarian" treatment withdrawal. No proponent seems to propose that treatments should be withheld from people who cannot be held (blame) responsible for their disease or injury. However, it is perfectly possible to maintain that a patient is responsible for having damaged her own health without also maintaining that any treatment should therefore be withheld from her. We should therefore distinguish between two conditions that both have to be satisfied for the "luck egalitarian" argument to be tenable: blame responsibility for a person's disease must be justifiably assignable to herself. Furthermore, this blame responsibility should be a sufficient moral reason for denying her access to healthcare resources that would otherwise have been available to her.

Responsibility is closely connected to causality, although that connection is mostly "merely implied" in the literature on individuals' responsibility for their own health (Guttman and Ressler 2001, 119, cf. Friesen 2018, 54). We normally only hold people blame responsible for what they have causally contributed to. Clearly, causality is not sufficient to justify blame; other conditions have to be satisfied, such as control over one's actions and knowledge of their causal implications (Driver 2008; Friesen 2018; Kelley 2005). However, for our present purpose, it is sufficient to note that causality is a necessary requirement. If a person did not causally contribute to her disease, then it would be difficult to claim that she should be held responsible for it. The examples used in the literature on "luck egalitarianism" all concern diseases that are taken to be caused by the patient's lifestyle or behavior.[4]

This gives rise to the three-tiered argument structure shown in Figure 6.1. The "luck egalitarian" argument for treatment withdrawal depends on three necessary conditions:

1. It is possible to determine whether a patient has caused her own disease or injury.
2. Blame responsibility can be justifiably assigned to the patient if she has caused her own disease or injury.
3. This blame responsibility is a sufficient moral reason for denying the patient some resources or treatments that would otherwise have been available to her.

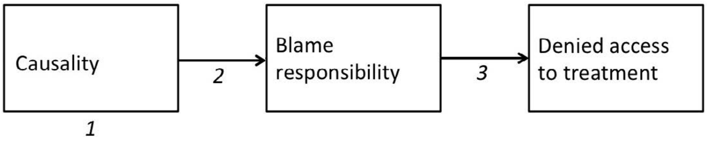

Figure 6.1 The three tiers of the justification of "luck egalitarianism"

The first of these conditions is epistemological and empirical, whereas the other two concern the validity of moral principles. We are now going to analyze each of the three conditions.

6.3 First Step: Causality

Causality is central in moral philosophy, not least due to the close connections between causality and the notion of consequences (of an action). It is usually assumed that causality is a well-defined and in principle well-understood phenomenon that provides us with a factual basis for our deliberations on value-laden concepts such as that of responsibility. Furthermore, causality as it appears in moral philosophy is usually conceived in the same way as in most everyday discussions, namely as consisting of (binary) relations between cause and effect. If asked to exemplify causality, most of us would bring up an example of such a relation between a cause and an effect. For instance, you turn the switch (the cause), and the lamp comes on (the effect). You hit three keys on the piano (the cause) and a B minor chord resonates in the room (the effect). An avalanche sweeps down the ski slope (the cause), and several skiers are buried in the snow (the effect). In these and many other cases, the cause–effect relationship adequately describes how different events in the world are connected with each other. But, this does not hold in general. Many processes in the world are not so well described in terms of cause–effect relations. This applies for instance to the movements of celestial bodies in the solar system. We can describe certain aspects of their movements in terms of cause–effect relationships; for instance, the movements of the moon can be said to cause tides. However, the whole pattern of movement in the solar system cannot be adequately described in terms of such cause–effect relationships. What goes on is a complex simultaneous interaction of mutual influences (Gómez, Masdemont, and Mondelo 2002). The same applies to other complex systems, such as ecosystems and the human body.

We have to distinguish, therefore, between two meanings of causality. In most everyday discussions, as well as deliberations in moral philosophy, the focus is on binary cause–effect relations. We can call this *CE-causality*. But, on other occasions, in particular in science, our talk of causality refers much more generally to the patterns of interdependencies that hold among objects and events in the world. How do events at some points in space-time restrict, or even determine, what happens at other points in space-time? When asking such questions, we are looking for the patterns of determination in the world, which we can call *world causality*. Importantly, world causality is a feature of the world, which we try to describe with the linguistic and mathematical means that are available to us. CE-causality is one of these means of description. In other words, CE-causality is a model of world causality. It is in many ways a successful and highly helpful model, but it has limitations, and as always, it is essential to distinguish between the real world and the models we use to describe it.

One of the problems with CE-causality is that it is *indeterminate*. By this I mean that for a given effect (E), there are often several causal factors that can plausibly be called "the cause" (C). This was clearly pointed out already by John Stuart Mill (Mill 1843/1996, 327–34). In the modern discussion on causality, this insight has often been exemplified by referring to the cause of cholera. If asked for the cause of that disease, a bacteriologist will tell you that it is caused by pathogenic strains of the bacterium *Vibrio cholerae*. However, an epidemiologist or public health expert confronted by the same question will usually have another answer: cholera is caused by poor sanitation, in particular lack of clean drinking water (Rizzi and Pedersen 1992). These different answers do not stem from conflicting opinions, only from differences in the focus that leads to selection of *the* cause among the causal factors contributing to the disease. Practically speaking, both answers are right.

It is not difficult to show that the indeterminateness of CE-causality creates severe problems for "luck egalitarianism." Let us consider two examples:

> Robert, who worked as a roofer, fell down from a roof and was severely hurt. According to his employer, the cause of the accident was that Robert lost his balance when carrying a large metal sheet that he should not have carried alone. His trade union representative says that the cause of the accident was that the employer had failed to implement adequate fall protection on the roof.

Both these accounts of what caused the accident may be right, in the same sense as the two explanations of cholera. There is no ground for claiming that one of them is objectively right and the other wrong. This is important, because if Robert is taken to a hospital honoring "luck egalitarian" principles, then the choice between these two causal descriptions can be decisive for what treatment he receives.

> Evelyn got severely ill from eating a stew that she cooked with mushrooms she had picked herself. Some say the cause of her getting ill is just that she picked the wrong mushrooms for her stew. However, after studying the mushroom field guide that she used, a mushroom expert says that the real cause was that this booklet presented an edible species without warning against a very similar, highly poisonous species.

The analysis of this case is parallel to that of the previous one.

Another serious problem with the application of the CE model to disease causality is that the causality of diseases is largely *stochastic*. People who are exposed to a risk factor for a particular disease will have an increased risk of contracting that disease, but the risk is usually not zero for those who have not been exposed to the risk factor in question. For instance, a sedentary way of life increases the risk of cardiac disease. However, some non-exercisers would have contracted the disease anyhow. Although we have reliable knowledge on the group level that a sedentary way of life increases the prevalence of cardiac disease, we have no means to transfer that certainty to the individual level. We cannot identify the persons who would have been spared from the disease if they had exercised more. The same applies to other connections between diseases and the way we live, including diseases connected with alcohol and smoking. This means that the so-called lifestyle diseases that are the common targets of "luck egalitarians" have no sure causal connection with lifestyles in the individual cases (contrary to the group level, on which many such connections are known beyond reasonable doubt).[5]

In summary, disease causality is usually both indeterminate and stochastic. This means that there is no objective or otherwise indisputable answer to the question of whether or not a person has caused her own disease or injury. Thus, condition (1) of Section 6.2 is not satisfied.

6.4 Second Step: From Causality to Blame Responsibility

Let us now turn to the second step in the justification of a "luck egalitarian" view of access to healthcare, namely the step from causality to blame responsibility. In this section, we will assume for the sake of argument that it can be objectively determined whether a patient has caused her own disease or injury. We will consider cases of what intuitively seem to be self-inflicted diseases or injuries and refrain from the type of critical analysis of these intuitive impressions performed in the previous section. Our first case is Bogdan, who has a severe traumatic brain injury that may require costly long-term treatment and care. Would we assign blame responsibility to him in the following cases?

1. He was injured when working as a paramedic in a war zone.
2. He was injured when working as a journalist in a war zone.
3. He was injured when visiting a war zone in order to collect background information for a novel.
4. He was injured by a falling beam when working as a fire-fighter.
5. He was injured by a falling beam when working as a carpenter.

6. He was injured by a falling bough when jogging in a forest.
7. He was a popular boxer. His head injury resulted from a long series of repeated concussions incurred during matches.

In all these cases, Bogdan can be said to have caused or at least causally contributed to his injury. He could have avoided the injury by a different choice, for instance (in the first case) by not working as a paramedic in a war zone, or (in the last case) by not becoming a boxer. However, in spite of his causal role, few of us would hold him blame responsible in all these cases.[6] For instance, most of us would presumably be unwilling to assign blame responsibility to the paramedic in the war zone (case 1). The reason for this is, of course, that the paramedic's activity is morally laudable.

More generally speaking, there are many unhealthy and dangerous activities that are regarded as desirable or at least acceptable, and we tend not to hold people responsible for negative consequences of engaging in such activities (Buyx 2008). This is not only specific to physical traumas but also applies to other types of medical conditions. For instance, a physician or nurse who volunteers to work with the treatment or prevention of a dangerous infection can run a considerable risk to contract the infection. If she contracts it, then she can reasonably be said to have contributed causally to her own disease. However, it can safely be assumed that few of us would be willing to assign blame responsibility to her.[7] In contrast, most of us would be much less hesitant to assign blame responsibility to those who engage in other types of activities that increase the risk of an infection, such as needle-sharing and unauthorized entry into a quarantine. Similar differences apply to many other types of behavior. For instance, promiscuity increases the risk of some diseases. These diseases are commonly taken to be self-inflicted, and they are connected with both blame and stigma. However, there are also at least three serious diseases (breast, ovarian, and uterine cancer) that are more common among sexually non-active than sexually active women. Therefore, nuns run an increased risk of these diseases (Britt and Short 2012). But (luckily) no one seems to assign blame responsibility to the unfortunate nuns who fall victim to one of these diseases.

All this adds up to the conclusion that there is no invariable or objective inference from a person's causing or causally contributing to her own disease to her being held blame responsible for it. Responsibility ascriptions depend on "not only the way in which an individual knowingly contributed to a negative health outcome, but also whether or not the individual engaged in a socially undesirable behavior" (Friesen 2018, 56). The major intervening factor, which determines whether causality leads to blame responsibility, is our moral appraisal of the action or activity in question.

However, "luck egalitarians" usually evade this problem by applying their principles only to a small selection of "option luck" activities that give rise to diseases, mainly overeating, unhealthy diets, lack of exercise, drinking, smoking, and drug use. As forcefully argued by Phoebe Friesen, this selection amounts to an "exclusive focus on highly stigmatized behaviors" (Friesen 2018, 56). Others have noted that the behaviors selected for blame responsibility largely coincide with ancient sins such as gluttony, sloth, and lust (Guttman and Ressler 2001, 118; Wikler 2002, 52). But for the sake of argument, let us accept this selective application. Will it solve the problem? In other words, can the inference from someone having caused her own disease to her being blame responsible for it be drawn without exceptions in the typical "lifestyle" cases that are at the center of the debate? Consider the following example:

> Arianna has always led an extremely sedentary way of life, with minimal exercise and high consumption of fatty junk food. She is obese, and she now has a severe coronary condition. According to the expert opinion of several cardiologists, she would not have acquired that condition if she had stayed slim and eaten healthy food.

This seems to be a prototypical case of a patient whose behavior caused her own disease. If Arianna did not cause her own disease, then who can be said to have done so? And shouldn't she be held blame responsible for her disease? Let us consider some alternative pieces of additional information that may – or may not – have an influence on our appraisal:

1. She has been held captive for the last 20 years in a cellar. Her captor provided her with an abundance of junk food but did not force her to eat more than she wanted. She had access to an exercise bicycle, but almost never used it.
2. In the last 20 years, she has worked long hours every day in a charity that helps victims of domestic violence. She has not had time for exercise, and she has saved time by ordering almost all her meals from a nearby fast food restaurant.
3. In the last 20 years, she has worked incessantly, day and night, as the CEO of a commercially successful pharmaceutical company whose products have saved thousands of lives. She has not had time for exercise, and late at night she would order copious amounts of food from a fast food restaurant.[8]
4. In the last 20 years, she has worked incessantly, day and night, as the CEO of a chemical company. The company's operations have always been legal, but they have often been criticized by environmental organizations. She has not had time for exercise, and late at night she would order copious amounts of food from a fast food restaurant.

It is difficult to believe that anyone would consider Arianna blameworthy or blame responsible for her bad health in all these cases. At the very minimum, we do not expect anyone to do so in the first case, in which she was captive. The important point is that by making exceptions like this, one accepts that there is no certain inference to be drawn from "X's overeating has caused her own disease" to "X is (blame) responsible for her own disease." Thus, even for the "lifestyle" behavior of eating too much, blame responsibility cannot always be justifiably assigned to persons who have, in the sense of option luck choices, caused their own disease. A moral appraisal will have to be made in each individual case before the person can reasonably be held blame responsible. Thus, condition (2) in Section 6.2 is not satisfied. (It is of course possible to mend "luck egalitarianism" by adding such moral appraisals. But who wants to live in a society that makes such moralizing appraisals of individuals' life choices?)

6.5 Third Step: From Blame Responsibility to Withdrawn Healthcare

In this section, we will assume for the sake of argument (but contrary to the results of Sections 6.3–6.4) that it can be clearly and non-arbitrarily determined whether a person has caused her disease or injury and that when she has, then she can reasonably be held responsible for it. The question is then: is it morally right to deprive her, for this reason, of healthcare that she would otherwise have received?[9] I will present four arguments, each of which is in itself sufficient to show that it is not morally right. This is the aspect of our topic that has received the most attention in previous literature, and parts of the argumentation summarize previous discussions.

1. The proposed deprivation of healthcare is a cruel and unusual punishment, administered without proper legal procedures.

Someone might protest that this is not a matter of punishment but of priority-setting. However, a legally instituted rule that denies a person certain medical treatments if she has committed certain actions will be a *de facto* punishment, and it will be perceived as such whatever the legislator chooses to call it.[10] Furthermore, denying someone healthcare is normally considered a cruel and unusual punishment (Rothschild 2019). In addition to being cruel and unusual, it would also in this case often be stochastic. For some of the affected persons, treatment withdrawal will not make any large difference, but for others it can have lethal consequences. Typically, it cannot be predicted who will be affected how.

Although "luck egalitarians" do not provide much detail on the decision-making process, they clearly assume that the decisions on treatment withdrawal will be made by healthcare providers or by insurers. Presumably, physicians would make the decisions. This would result in "punishment without a hearing or trial by individuals who were effectively jury, judge, and executioner rolled into one" (Harris 1995). Physicians do not have education for any of these roles, and it is difficult to see how a person with this combination of roles could uphold the patient's right to be presumed innocent (i.e., presumed not to have caused her own disease, unless and until it has been established beyond a reasonable doubt that she has done so) (Clavien and Hurst 2020, 188; Wikler 2002, 53).

2. It is unfair to people who have done nothing worse than mismanaging their own health that they are denied treatment that is available to people who have done much worse things.

There are quite a few "blameworthy individuals including convicted criminals that have committed much more heinous crimes than repeatedly picking the wrong dessert" (Clavien and Hurst 2020, 189). "Luck egalitarianism" does not imply that their access to healthcare should be reduced. To be concrete, compare the following two persons. They both have a severe heart condition and would gain from an expensive treatment:

> Cynthia has worked many years as a nursing assistant, and her work is much appreciated by patients, nurses and doctors. However, she has mismanaged her personal health by smoking several packages of cigarettes every day. According to several cardiologists, her heart condition is the result of her smoking.
>
> Donald is the CEO of a tobacco company. He has energetically and successfully promoted the sale of cigarettes on new markets in low- and middle-income countries. Thousands of people have died from the products marketed by the company he leads. However, himself he has never smoked. His diet is healthy, and he drinks very little alcohol. He also has sound exercise routines. According to several cardiologists, he got the disease in spite of an exemplary way of life.

Contrary to Donald, Cynthia would certainly be a candidate for "luck egalitarian" restrictions in access to health. However, we may well ask why the victims of the tobacco epidemic should be penalized, and not its instigators. In the absence of a plausible answer to this question, it is difficult to avoid the conclusion that "luck egalitarianism" is a victim-blaming ideology.

3. A system that restricts healthcare access for self-inflicted diseases will have a negative impact on the patient–physician relationship, with potential negative consequences for the quality of healthcare.

The patient–physician relationship is in essential respects fiduciary, which means that it is a relationship of trust, based on the physician's undertaking to act in the interest of the patient in the matters covered by the relationship – namely the patient's health (Bartlett 1997; Bester

2020; Faunce and Bolsin 2005). One important component of this relationship is that it is in the patient's interest to answer the physician's questions truthfully. This is important since physicians need that information to be able to make the right recommendations. For instance, there is usually no other way than interviewing the patient to find out how much she exercises, how long she has smoked, whether she has used drugs, what type of food she eats, or whether she is at particular risk of sexually transmitted diseases. Such information can be important for the choice of treatment. All this would be changed with the introduction of a system that withholds treatment from patients who are deemed to be responsible for their own disease or injury (Ho 2008, 81; cf. Waller 2005, 186). Even if physicians will not be the decision-makers who deny treatment, information from their examination of the patient will be crucial for the decision. A patient could for instance try to make herself eligible for an expensive treatment by claiming to have started to smoke just a few months ago. In a system of blame-based treatment denial, we can expect information to be available on the Internet – for free or for sale – on what one should withhold from one's doctor in order not to be denied any treatment.

4. Since the rich can always buy the care they need, a system that restricts healthcare access for self-inflicted diseases will in practice not affect them. Only the poor, not the rich, will be punished for unhealthy behavior.

It is in practice not feasible to prevent private organizations from offering medical treatment against money. Even in the unlikely case of a prohibition to offer certain treatments outside of the public health system, people in need of treatment could go abroad for it. However, no "luck egalitarian" seems to have proposed that rich people should be prevented from buying all the healthcare they can afford.[11] Therefore, blame-based treatment denial will selectively affect the poor (Huzum 2009, 206–7; Persaud 1995, 284). In its practical implementation, "luck egalitarianism" will not, as is usually claimed, withdraw healthcare from people with diseases classified as self-inflicted. It will withdraw healthcare from *poor* people with such diseases.

6.6 Conclusion

We began by identifying three conditions that must be satisfied for "luck egalitarianism" to be a workable standpoint at all: (1) It must be possible to determine whether a person caused her own disease or injury, (2) it must be justifiable to hold a person (blame) responsible for her disease or injury if she caused it herself, and (3) this blame responsibility is a sufficient moral reason to withhold some beneficial treatments from the person, which she would otherwise have been offered.

The first condition fails due to the indeterminate nature of our common notion of causality and to the stochastic nature of disease causation. The second condition fails since it would have morally absurd consequences to hold individuals blame responsible for the negative effects on their health of morally laudable or at least acceptable (option luck) choices that they have made. In practice, "luck egalitarians" try to evade this problem by only applying their principles to certain behaviors such as smoking and overeating that are believed always to be blameworthy, but we have shown that the problem can arise in such cases as well. The third condition fails since there are several strong moral reasons not to deprive people of healthcare who have (presumably) caused their own disease. Such deprivation would be a cruel and unusual punishment with potentially lethal consequences, but administered without procedures ensuring basic legal principles such as the presumption of innocence. While people who have mismanaged their own health would be deprived of healthcare, people who have done much worse things, such as damaging the health of others, are not affected. Furthermore, such a system would harm the patient–physician relationship by making it dangerous for patients to give truthful answers to the doctor's questions about their current and previous ways of life.

Perhaps most importantly, the rich will always be able to buy the healthcare they need. A "luck egalitarian" system for the withdrawal of healthcare resources would leave the privileged unaffected while punishing the poor for behaviors into which they were deceived by the marketing tricks of powerful pathogenic companies such as the tobacco, alcohol, and soft drink industries. Such a policy does not answer to any reasonable definition of egalitarianism. This is the reason why I use scare quotes around the phrase "luck egalitarianism." It is usually preferable to use the terms introduced by the initiators of an idea or standpoint, but there is also a limit to how misleading terms one should use for a concept. An anti-egalitarian policy should not be called egalitarian. "Luck anti-egalitarianism" is a more suitable term for the standpoint that has been promoted under the name "luck egalitarianism."

Notes

[1] The concept of brute luck combines misfortunes caused by nature and misfortunes caused by other individuals or by social arrangements into one and the same category. This is problematic, since it can induce us to treat changeable social conditions as a matter of luck rather than of social reform. For instance, being born as a woman in a misogynist society or as a person of color in a white supremacist society should not be treated as a matter of brute luck but as a matter of brute oppression.

[2] The terms "task responsibility" and "blame responsibility" were introduced by Goodin (1987, 167–8). With a common but less adequate terminology, blame responsibility is called "backward-looking responsibility" and task responsibility is called "forward-looking responsibility."

[3] For a more detailed discussion of the dissociation between ascriptions of blame and task responsibility to patients, see Hansson (2018). Many other authors have indicated that patients should be entrusted with task responsibilities but not unnecessarily burdened with blame responsibility (Dougherty 1993, 118; Feiring 2008; Gonzalez, Goeppinger, and Lorig 1990, 137; Kelley 2005; Pickard 2017; Waller 2005). On the negative effects of blaming patients, see also Doley et al. (2017), Jerpseth et al. (2021), Marantz (1990), and Talbot and Branley-Bell (2021).

[4] The term "lifestyle" is standard terminology in the "luck egalitarian" literature. However, the word "style" has the disadvantage of giving the impression of something freely chosen that can be changed with the same ease as affluent people can change their "clothes style." Much of what is commonly called "lifestyle" is in fact determined by social conditions and by the environment in which the individual lives (Watt 2007). Terms such as "living conditions" or "way of life" would be preferable.

[5] Roemer (1995) proposed what he called "a way to decide which aspects of a person's behavior are due to circumstance and which to autonomous choice." For smoking-related diseases, he proposed that people should be divided into "types" according to characteristics correlated with tobacco use, such as age, ethnicity, gender, and occupation. Smokers will then be held responsible for their smoking according to how much they have smoked, compared to others of their type. Thus, a 60-year-old white female college professor who is at the 80th percentile of the smoking distribution among persons of her type will receive "socially financed medical care" for smoking-related lung cancer to the same extent as a 30-year-old Black male steelworker who is at the 80th percentile within his type. This attempt to derive individual causality or responsibility from correlations on the group level does not have much credibility. As noted by Norman Daniels, it implies that "[i]f skiing is a common behavior of the rich but not of poor working people, then the poor skier is more responsible for his or her broken leg in a skiing accident than the rich skier" (Daniels 2002, 254; cf. Hurley 2002, 51–4).

[6] The problem of responsibility attribution to different kinds of physical injuries has also been discussed by Anderson (1999) and Huzum (2009) among others.

[7] Albertsen and Thaysen (2017) attempt to save "luck egalitarianism" from the conclusion that such a person's claim to medical treatment is weakened by her voluntary risk-taking. Their solution is to apply "luck egalitarian" withdrawal of healthcare only to persons who choose to expose themselves to a danger that they have themselves created, not to persons who choose to expose themselves to a danger they have not created. However, this criterion does not seem to capture the intuitions that it is intended to reflect. For instance, a person who enters an Ebola quarantine in order to work there as a nurse and a person who enters the same quarantine in order to steal medical equipment have both chosen to expose themselves to a danger they have not created. This seems to be rather irrelevant for the moral appraisal of their respective actions. Their intentions seem much more relevant. The criterion also has the consequence that a person creating a danger to which she exposes herself can lose rights to healthcare, whereas someone who creates dangers to which she only exposes others runs no such risk (see Section 6.5).

[8] This is essentially a case described by Waller (2005, 181).

[9] Segall tries to escape the morally repulsive consequences of "luck egalitarianism" by prescribing that it has to be combined with some other moral principle, such as "the moral requirement to meet basic needs, including basic medical needs," which would ensure that patients with option luck diseases receive treatment (Segall 2010, 68). He does not clarify the meaning of "basic" in "basic medical needs." The crucial question here, which he does not answer, is as follows: does the "basic medical needs" principle ensure that everyone who has a medical condition due to her own option luck choices receives the same treatment as someone who has the same condition due to brute luck? If the answer is no, then the ethical criticism against "luck egalitarianism" still stands. On the other hand, if the answer is yes, then "luck egalitarianism" appears to be redundant in terms of action-guiding implications, and Occam's razor should be applied.

[10] The term "punishment" for such deprivation of healthcare has frequently been used in the literature (Brown and Savulescu 2021; Fleurbaey 1995, 40–1; Harris 1995; Huzum 2009, 198). Voigt (2007) claims that the term "punishment" is inadequate since luck egalitarians "are not involved in moral evaluations when deciding whether or not a given inequality is just" (p. 392). However, he also says: "To draw the distinction between option luck and brute luck we must make certain normative judgements about what we can reasonably expect from agents" (p. 397).

[11] Roemer (1995) explicitly restricted the withdrawal of medical resources to "socially financed medical care."

References

Albertsen, Andreas, and Jens Damgaard Thaysen. 2017. "Distributive Justice and the Harm to Medical Professionals Fighting Epidemics." *Journal of Medical Ethics* 43: 861–4.

Anderson, Elizabeth S. 1999. "What Is the Point of Equality?" *Ethics* 109 (2): 287–337.

Arneson, Richard J. 2000. "Luck Egalitarianism and Prioritarianism." *Ethics* 110 (2): 339–49.

Bartlett, Peter. 1997. "Doctors as Fiduciaries: Equitable Regulation of the Doctor-Patient Relationship." *Medical Law Review* 5: 193–224.

Bester, Johan Christiaan. 2020. "Defensive Practice Is Indefensible: How Defensive Medicine Runs Counter to the Ethical and Professional Obligations of Clinicians." *Medicine, Health Care and Philosophy* 23 (3): 413–20.

Britt, Kara, and Roger Short. 2012. "The Plight of Nuns: Hazards of Nulliparity." *Lancet* 379 (9834): 2322–3.

Brown, Rebecca C.H., and Julian Savulescu. 2021. "Personal Responsibility for Cardiac Health: What Are the Ethical Demands?" *Heart* 107 (9): 765–6.

Buyx, Alena M. 2008. "Personal Responsibility for Health As a Rationing Criterion: Why We Don't Like It and Why Maybe We Should." *Journal of Medical Ethics* 34 (12): 871–4.

Clavien, Christine, and Samia Hurst. 2020. "The Undeserving Sick? An Evaluation of Patients' Responsibility for Their Health Condition." *Cambridge Quarterly of Healthcare Ethics* 29 (2): 175–91.

Cohen, Gerald A. 1989. "On the Currency of Egalitarian Justice." *Ethics* 99 (4): 906–44.

Daniels, Norman. 2002. "Democratic Equality. Rawls's Complex Egalitarianism." In *The Cambridge Companion to Rawls*, edited by Samuel Freeman, 241–76. Cambridge: Cambridge University Press.

Doley, Joanna R., Laura M. Hart, Arthur A. Stukas, Katja Petrovic, Ayoub Bouguettaya, and Susan J. Paxton. 2017. "Interventions to Reduce the Stigma of Eating Disorders: A Systematic Review and Meta-Analysis." *International Journal of Eating Disorders* 50 (3): 210–30.

Dougherty, Charles J. 1993. "Bad Faith and Victim-Blaming: The Limits of Health Promotion." *Health Care Analysis* 1: 111–9.

Driver, Julia. 2008. "Attributions of Causation and Moral Responsibility." In *Moral Psychology*, *Volume 2*: *The Cognitive Science of Morality*, edited by Walter Sinnott-Armstrong, 423–39. Cambridge: MIT Press.

Dworkin, Ronald. 1981. "What Is Equality? Part 2: Equality of Resources." *Philosophy and Public Affairs* 10 (4): 283–345.

Faunce, Thomas Alured, and Stephen N. C. Bolsin. 2005. "Fiduciary Disclosure of Medical Mistakes: The Duty to Promptly Notify Patients of Adverse Health Care Events." *Journal of Law and Medicine* 12: 478–82.

Feiring, Eli. 2008. "Lifestyle, Responsibility and Justice." *Journal of Medical Ethics* 34 (1): 33–6.

Fierlbeck, Katherine. 1996. "Policy and Ideology: The Politics of Post-Reform Health Policy in the United Kingdom." *International Journal of Health Services* 26 (3): 529–46.

Fleurbaey, Marc. 1995. "Equal Opportunity or Equal Social Outcome?" *Economics and Philosophy* 11 (1): 25–55.

Friesen, Phoebe. 2018. "Personal Responsibility within Health Policy: Unethical and Ineffective." *Journal of Medical Ethics* 44 (1): 53–8.

Glannon, Walter. 1998. "Responsibility, Alcoholism and Liver Transplantation." *Journal of Medicine and Philosophy* 23: 31–49.

Gómez, Gerard, Josep J. Masdemont, and Josep-Maria Mondelo. 2002. "Solar System Models with a Selected Set of Frequencies." *Astronomy and Astrophysics* 390 (2): 733–49.

Gonzalez, Virginia M., Jean Goeppinger, and Kate Lorig. 1990. "Four Psychosocial Theories and Their Application to Patient Education and Clinical Practice." *Arthritis Care and Research* 3 (3): 132–43.

Goodin, Robert E. 1987. "Apportioning Responsibilities." *Law and Philosophy* 6: 167–85.

Guttman, Nurit, and William Harris Ressler. 2001. "On Being Responsible: Ethical Issues in Appeals to Personal Responsibility in Health Campaigns." *Journal of Health Communication: International Perspectives* 6: 117–36.

Hansson, Sven Ove. 2018. "The Ethics of Making Patients Responsible." *Cambridge Quarterly of Healthcare Ethics* 27: 87–92.

Harris, John. 1995. "Could We Hold People Responsible for Their Own Adverse Health?" *Journal of Contemporary Health Law and Policy* 12 (1): 147–53.

Ho, Dien. 2008. "When Good Organs Go to Bad People." *Bioethics* 22 (2): 77–83.

Hurley, Susan. 2002. “Roemer on Responsibility and Equality.” *Law and Philosophy* 21 (1): 39–64.

Huzum, Eugen. 2009. “The Principle of Responsibility for Illness and Its Application in the Allocation of Health Care A Critical Analysis.” In *Autonomy, Responsibility, and Health Care*, *Critical Reflections*, edited by Bogdan Olaru, 191–218. Bucharest: Zeta books.

Jerpseth, Heidi, Ingrid R. Knutsen, Kari T. Jensen, and Kristin Halvorsen. 2021. “Mirror of Shame: Patients Experiences of Late-Stage COPD. A Qualitative Study.” *Journal of Clinical Nursing* 30: 2854–62.

Kelley, Maureen. 2005. “Limits on Patient Responsibility.” *The Journal of Medicine and Philosophy* 30 (2): 189–206.

Marantz, Paul R. 1990. “Blaming the Victim: The Negative Consequences of Preventive Medicine.” *American Journal of Public Health* 80 (10): 1186–7.

Mill, John Stuart. 1843/1996. *A System of Logic*. *Collected Works*, vol. 7. Toronto: University of Toronto Press.

Persaud, Rajendra. 1995. “Smoker’s Rights to Health Care.” *Journal of Medical Ethics* 21: 281–7.

Persson, Karl. 2013. “The Right Perspective on Responsibility for Ill Health.” *Medicine, Health Care and Philosophy* 16 (3): 429–41.

Pickard, Hanna. 2017. “Responsibility without Blame for Addiction.” *Neuroethics* 10 (1): 169–80.

Rakowski, Eric. 1991. *Equal Justice*. New York: Oxford University Press.

Rizzi, Dominick A., and Stig Andur Pedersen. 1992. “Causality in Medicine: Towards a Theory and Terminology.” *Theoretical Medicine* 13: 233–54.

Roemer, John. 1995. Equality and Responsibility. *Boston Review* 20. https://www.bostonreview.net/forum/equality-and-responsibility/

Rothschild, Molly. 2019. “Cruel and Unusual Prison Healthcare: A Look at the Arizona Class Action Litigation of Parsons v. Ryan and Systemic Deficiencies of Private Health Services in Prison.” *Arizona Law Review* 61: 945–81.

Segall, Shlomi. 2010. *Health, Luck and Justice*. Princeton: Princeton University Press.

Talbot, Catherine V, and Dawn Branley-Bell. 2022. “#BetterHealth: A Qualitative Analysis of Reactions to the UK Government’s Better Health Campaign.” *Journal of Health Psychology* 27 (5): 1252–58.

Underwood, Malcolm J., and John S. Bailey. 1993. “Coronary Bypass Surgery Should Not Be Offered to Smokers.” *BMJ* 306: 1047–50.

Voigt, Kristin. 2007. “The Harshness Objection: Is Luck Egalitarianism Too Harsh on the Victims of Option Luck?” *Ethical Theory and Moral Practice* 10 (4): 389–407.

Waller, Bruce N. 2005. “Responsibility and Health.” *Cambridge Quarterly of Healthcare Ethics* 14: 177–88.

Watt, Richard Geddie. 2007. “From Victim Blaming to Upstream Action: Tackling the Social Determinants of Oral Health Inequalities.” *Community Dentistry and Oral Epidemiology* 35: 1–11.

Wikler, Daniel. 2002. “Personal and Social Responsibility for Health.” *Ethics and International Affairs* 16: 47–55.

7 Moral Responsibility and Public Health Risks

Examples from the Coronavirus Pandemic

Jessica Nihlén Fahlquist

7.1 Introduction

Many societal problems require both individual and collective actions. Some of these problems entail personal as well as collective responsibility. Among the most urgent problems of our time are those related to the environment and public health. That public health is a great challenge became even clearer during the coronavirus pandemic that broke out in 2020.[1] The pandemic brought ethical questions to the fore. Although ethical questions about public health had been discussed before, they became part of public debate, engaging laypeople as well as experts to an unprecedented extent.

This chapter focuses on public health risks and moral responsibility, with examples from the coronavirus pandemic. In order to analyze and discuss the pandemic from the perspective of risk and responsibility, it builds a conceptual framework focusing on elements such as public health, risk ethics, virtue, individual and collective responsibility, fairness and efficacy, as well as backward- and forward-looking responsibility. As this chapter emphasizes, risk and responsibility are closely connected, and they are both salient in modern society, in which substantial attention is directed toward managing and communicating risks.

The ethical questions at the heart of this chapter primarily concern matters of moral responsibility and risk that arose during the coronavirus pandemic. These ethical questions, however, prove to be relevant to pandemics and even infectious diseases generally. The questions revolve around matters of governmental and individual responsibility for collective problems, as well as the challenges associated with ascribing responsibility for, among other things, risk management.

As discussed below, both governments and individuals need to take responsibility for these sorts of collective problems for reasons relating to both efficacy and fairness. Furthermore, governments must communicate clearly (a) how they balance conflicts between collective health and individual rights and values and (b) what the chosen strategy entails in terms of

DOI: 10.4324/9781003276029-10

collective and individual responsibility. These tasks require an open public discourse about the values involved. While experts can provide numbers and facts, individuals need to be involved in determining which decisions are made and how decisions are made. Success requires attention to ethical values from all involved. Individuals need to cultivate character traits that can help manage this pandemic and prevent new ones (e.g., compassion). Governments must facilitate the development of such character traits by building trust and solidarity with and among citizens.

The chapter is structured as follows. Section 7.1 introduces insights from the area of risk ethics and discusses their relevance to public health. Section 7.2 provides a conceptual discussion of moral responsibility, focusing on various subtopics such as fairness and efficacy, backward- and forward-looking responsibility, as well as individual and collective responsibility. Section 7.3 uses the framework from the preceding sections and discusses the coronavirus pandemic from the perspective of individual and governmental responsibility. This is buttressed by analysis of topics, such as beneficence and justice, autonomy and non-maleficence, and trust and solidarity.

7.2 Risk Ethics and Public Health

Conventional risk-benefit analysis, management, and communication do not always include ethical values unless they can be included in a numerical calculation. A risk is considered acceptable if the benefits outweigh the risks. Although there are inevitable challenges when deciding on what counts as risks and benefits, this is a fairly straightforward method. However, although it is sometimes referred to as a value-neutral method, it is based on utilitarianism. Thus, it does not necessarily cover all important values but is limited to what counts as risk and benefit. Research shows that laypeople consider a wide range of other aspects. For example, laypeople care about voluntariness in risk-taking and justice in risk-benefit distribution. They do not necessarily consider a low probability to be more important than a potentially catastrophic consequence (e.g., Roeser et al. 2012; Slovic 2000). It has been argued that effective risk communication needs to consider these other aspects of risks (Árvai and Rivers 2014).

Although some of the aspects that laypeople take into consideration may be considered irrational, too emotional, or even unethical, values like justice and voluntariness are normatively important (e.g., Roeser 2017). Hence, not only *do* people care about fairness in a risk-benefit distribution, we *should* care about it (Roeser 2017). Justice and fairness are not merely a matter of efficacy, but of ethical legitimacy and justifiability. If a decision concerning a government intervention benefits wealthy people, but poor people are exposed to risks, this is an ethically problematic situation

(Hermansson and Hansson 2007). Similarly, if a decision benefits adults, but leads to an exposure of risks to children, it is ethically questionable. In these cases, a collectivist risk-weighing method could lead to a conclusion that such risks are acceptable because the benefits outweigh the risks from a societal perspective. However, taking voluntariness and justice into account and how they affect individuals and minorities, they may not be considered acceptable. The collectivist risk-weighing principle underlying conventional risk analysis largely ignores the fact that risks are not "free-floating entities," but that individuals take risks or are exposed to them (Hansson 2007; Hermansson and Hansson 2007).

When faced with ethical problems, philosophers commonly apply ethical theories. However, as Hansson (2003) and others argue, ethical theories are not easily applicable to problems of risk. Whereas ethical theories deal with certainty and, to some extent, presuppose a deterministic world, the actual world is more complex and less certain (Hansson 2003). A problem with utilitarianism is that many aspects and values cannot simply be reduced to assignments of utility. However, duty- and rights-based theories would, strictly, not allow any risk of harm regardless of probabilities. If individuals have a right not to be harmed, this right is absolute and unconditional and everybody has a duty to respect the rights of others not to be harmed. This notion is based on a relatively simplistic conception of harm as binary: you are either harmed or not harmed by someone's actions. This would not be realistic in modern societies, characterized by risk (Hansson 2003). As the other authors in this book, I believe a fruitful approach to risk problems focuses on concepts and notions of moral responsibility (Nihlén Fahlquist 2018). This is partly due to the complexity of modern societies and partly because of the conceptual connection between risk and responsibility. I will return to this, but in order to tie this discussion to the coronavirus pandemic, I will first present the area of public health ethics.

Risk ethics has largely been developed in the context of technology and engineering (e.g., Asveld and Roeser 2009; Hansson 2013; Roeser et al. 2012). However, it is also highly relevant in the context of public health.

Just like risk ethics is a relatively new branch of ethics, public health ethics has become an area of research during recent decades. Most bioethical literature refers to the principles of biomedical ethics (Beauchamp and Childress 1994).

Beauchamp and Childress' idea was to find principles that could guide healthcare and that most people would agree on regardless of preferred ethical theory or perspective. They argue that the principles of beneficence, non-maleficence, autonomy, and justice are the most important tools for analyzing ethical issues in healthcare. Scholars interested in bioethics cannot avoid the importance of these principles even though substantial critique has been directed against their framework, and they have to be interpreted in actual cases.

During the last decades and along with the development of welfare states, the public health perspective has become more important. Public health policy and activities differ from regular healthcare. First, public health policy is aimed at the population at large and not individual patients. Second, it is collective, requiring government action. Third, it focuses on prevention rather than treatments (Dawson 2011). Public health aims to prevent disease, prolong life, and promote health (Royo-Bordonada and Román-Maestre 2015). The World Health Organization has a very ambitious definition of health, as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (World Health Organization 2022). Consequently, governments all over the world have a demanding agenda regarding risks that should be managed and communicated.

Aiming to prevent disease and promote health, public health is necessarily and primarily focused on analyzing, reducing, managing, and communicating risks. Because public health aims to improve the health of the population, it is not targeted at specific individuals. This feature is an expression of its utilitarian foundation and collectivist risk-weighing principle.

This focus requires a somewhat different way of approaching ethical aspects compared to regular medical healthcare. The principles and discussions emerging from that framework are inadequate. This is not to say they are inapplicable or irrelevant. The principles are relevant to public health.

For example, public health necessarily gives rise to potential conflicts between autonomy and beneficence. Aiming at maximizing health, as defined by governments, for the population at large, is not likely to be completely consistent with the interests and values of all individuals. One good example is the so-called lifestyle-related diseases that are increasing in many societies. Setting the goal to reduce obesity, for example, means that at some point people should lose weight. For obvious reasons, the weight of the population at large cannot be reduced without individuals doing some kind of work, increasing exercise, and decreasing food intake. If society sends messages in pursuit of this goal, there is a risk that obese individuals may be stigmatized and that individuals who do not conform are viewed as a burden (cf. Guttman and Salmon 2004). Furthermore, some people may value quality of life more than quantity of life, and the former may involve more or less unhealthy food and drink. Prohibiting smoking means that some people may feel that their current view of quality of life is threatened. Another example is when governments communicate messages to pregnant women and parents.

The message that all mothers should breastfeed their infants potentially has negative consequences for individual women who may feel inadequate when they have problems with breastfeeding (Nihlén Fahlquist 2016; Nihlén Fahlquist and Roeser 2011). Similarly, mandatory vaccination infringes on people's right to make decisions concerning their own health. Some public health measures are justified regardless of effects like these, but they should be ethically deliberated and followed up on as opposed to being routinely initiated. When there is evidence of a certain risk, it is not self-evident that it should be communicated to the public. Thus, conflicts between beneficence and individual autonomy are common in public health. However, there are other aspects that need to be included in public health ethics, which have evolved alongside the development of public health policy.

The example of lifestyle-related diseases has also been discussed in terms of responsibility distribution. Using the concept of "lifestyle" is partly based on an assumption that individuals choose their lifestyle from an array of lifestyles. This is particularly important given research about the socioeconomic determinants of health (e.g., Braveman and Gottlieb 2014).

The question is to what extent is health an individual responsibility and to what extent it is the responsibility of the government (e.g., Wikler 2002). In the following, I will discuss important distinctions and notions of moral responsibility that are useful in analyses of public health. I will then discuss the case of the coronavirus pandemic from the perspective of moral responsibility.

7.3 Moral Responsibility

As discussed above, applying ethical theories and bioethical principles to decisions concerning risks and benefits is complicated. A more promising approach is to put risks into a conceptual and normative framework focusing on moral responsibility.

One important reason for doing this is the conceptual interconnectedness between risk and responsibility. Human beings have experienced risks in all times, but they have traditionally been described as "dangers" or "threats". In modern society, these threats started to be conceptualized as "risks". This transition meant that dangers started to be seen as more or less manageable and controllable. Human beings and societies could and should be able to do something about them for protection. Thus, risks are now seen as calculable threats (See Beck 1992; Giddens 1999; Nihlén Fahlquist 2018).

By calculating probabilities, threats are no longer merely unknown or uncertain, but they are something that human beings can affect. This, in a sense, implies that someone is responsible for the consequences. Both concepts, risk and responsibility, are linked to control, action-taking, and decision-making. However, responsibility is a multifaceted and complex concept (Nihlén Fahlquist 2018; Van de Poel and Nihlén Fahlquist 2012).

There are three important distinctions that are useful to this analysis. First, responsibility can be ascribed for reasons of fairness or efficacy. Second, responsibility ascriptions can be backward-looking or forward-looking. Third, risk problems could be a matter of individual or collective responsibility, or a combination of these.

7.3.1 Fairness and Efficacy

Responsibility could be ascribed and distributed for reasons of fairness and/or efficacy. In public debates, these different aims and reasons are commonly controversial and sometimes conflated. Whereas someone may refer to fairness, someone else may be interested in the efficacy of a certain responsibility ascription. What is fair is not necessarily effective and vice versa. The question of fairness partly has to do with causation and wrongdoing (i.e., when we think that someone caused something and the action and its consequences involved harm and/or wrongdoing, it is considered fair to hold her responsible). However, people often disagree about causation as well as wrongdoing (Nihlén Fahlquist 2018). For example, it could be argued that a road crash was caused by the driver or the deficient road. Similarly, antibiotic resistance could be seen as being caused by inadequate governmental action and legislation, excessive prescriptions of medical doctors, or non-compliance of patients.

In contrast, responsibility could be ascribed for reasons of efficacy and efficiency. For example, when discussing antibiotic resistance or climate change, we may refer to what would be the most effective way to reduce these risks (i.e., whether individuals, the industry, or the government are primarily responsible). When using the concept of responsibility in this way, focus is on what should be done and by whom (Nihlén Fahlquist 2018).

The optimal responsibility distribution is both fair and effective, but conflicts of value may arise. Some theories favor fairness and others focus on efficacy. From a merit-based perspective, responsibility should be ascribed to the one who deserves to be held accountable because, for example, she presumably intended to do it and had the relevant knowledge. The idea is that an agent who knowingly and intentionally caused wrong should be held accountable. In contrast, consequentialist theories merely focus on the outcomes, and a utilitarian perspective entails the idea that responsibility should be ascribed in such a way that utility is maximized. If there is no expected utility, we should refrain from ascribing and distributing responsibility (Nihlén Fahlquist 2018). Thus, there are two different norms at stake when responsibility is to be distributed:

1. Deontological norm: A responsibility ascription or distribution ought to be fair; that is, based on merit and that which is deserved.
2. Consequentialist norm: A responsibility ascription or distribution ought to be effective; that is, have the best possible consequences compared to other responsibility ascriptions or distributions (Nihlén Fahlquist 2018).

7.3.2 Backward-Looking and Forward-Looking Responsibility

In addition to the two norms described above, responsibility could be backward-looking or forward-looking. When it comes to the first, responsibility ascriptions are commonly based on certain conditions. In ordinary language as well as in the philosophical debate, a basic notion is that a person is responsible if certain conditions are fulfilled. For example, if someone was forced to do something that we normally consider wrong, we are generally reluctant to hold her responsible (i.e., we would excuse her for acting wrongfully). van de Poel et al. (2011) argue that the following is a list of conditions that are often seen as prerequisites for holding an agent morally responsible in the sense of blameworthiness (van de Poel et al. 2011): (a) capacity, (b) causality, (c) knowledge, (d) freedom, and (e) wrongdoing.

In my view, there are primarily two notions of forward-looking responsibility that are particularly important in the context of risk management in society: responsibility as task or role, and responsibility as virtue. As individuals and professionals, we all have tasks every day, more or less formalized. They are often connected to our different roles. As a parent, a friend, or as a doctor or engineer, we have different tasks we are expected to take on.

This kind of responsibility is directed toward the future. In a policymaking context, the relevant tasks usually relate to goal-setting and goal-achievement—for example, concerning risk reduction and risk management. It could, for example, be a task to manage a pandemic, to reduce climate change, or minimize the risk of nuclear accidents. (Nihlén Fahlquist 2018).

Tasks and roles are relatively straightforward. Responsibility as a virtue is more complex, which is a strength as well as a weakness. It is harder to pinpoint, but it is both important and useful in a complex modern society characterized by risk problems. Responsibility as a virtue involves a notion that it is important to actively take responsibility and to develop and cultivate certain character traits and ways of acting and being in the world

(Nihlén Fahlquist 2018). Being a responsible person, in this sense, requires commitment, initiative, and judgment (Williams 2008). It entails a "willingness to make sacrifices in order to get involved" (Van Hooft 2006). According to Williams, it represents a "readiness to respond to a plurality of normative demands" (Williams 2008). Furthermore, the agent must be able to be trusted to exercise some degree of discretion (Nihlén Fahlquist 2018; Williams 2008).

I have suggested elsewhere that an agent is a responsible person, in the virtue-ethical sense, if

1. the agent *cares* about other people and the way the activities in which she partakes potentially affect other people (and the environment);
2. the agent has the emotional ability to *morally imagine* what those effects could be like and what risks might be involved in those activities; and
3. the agent has the cognitive ability to transform these concerns into practice and actions, i.e., she has *practical wisdom* (Nihlén Fahlquist 2018).

I will return to the notion of responsibility as a virtue in the section discussing the coronavirus pandemic.

7.3.3 Individual and Collective Responsibility

The third distinction that is useful in the context of risk problems, like public health, is individual and collective responsibility. Typically, societal problems require both individuals and governments to act.

Moral responsibility has traditionally been conceived as connecting to individual moral agents. The notion of collective moral responsibility has been considered with suspicion. This is primarily due to the historical experiences during the 20th century. For example, the German people were collectively blamed for atrocities committed during World War II. In relation to this, Lewis argues that it is dangerous to treat collectives like nations, families, or tribes as the main unit. He argues that such ideas are "simply the obverse of the tendency to set some abstract good of the community above the wellbeing of its individual members", connected to oppressive and totalitarian ideas and practices (Lewis 1948). Velasquez argues that ascriptions of collective responsibility are always reducible to individual responsibility (Velasquez 1983/1991). In contrast, Cooper argues that there is irreducible collective responsibility. One reason for this is that we do ascribe responsibility to collectives. Furthermore, he argues, it is impossible to make conclusions concerning individual blame responsibility from arguments concerning collective responsibility (Cooper 1968/1991).

Bovens argues that because complex organizations have extensive power to change people's lives, and we are dependent on them, collective responsibility is necessary (Bovens 1998). The discussion is ongoing.

In the context of societal risk management and communication, it is unrealistic to conceive activities and consequences for people and the environment merely in terms of individuals. In modern complex societies, it is unrealistic and arguably unethical, to avoid notions of collective responsibility. There is so much activity that takes place in organizations and the outcomes affect people and societies in various ways. In some cases, negative outcomes can be reduced to individual actions and wrongdoing, but not always. Technological risks, as well as benefits, are created in the context of collective undertakings. Climate change is caused both by individual actions and government and industry action and inaction. Similarly, public health problems require collective responsibility. It aims to prevent disease and improve the health of the population through state interventions. Therefore, both the agent and the target are collective entities. Public health activities are normally conducted by and through government agencies. On the other hand, governments and agencies consist of people, and it is important to keep that in mind. I have argued that public health professionals should develop the virtues of responsibility, compassion, and humility (Nihlén Fahlquist 2019). One of the main reasons for that is the power asymmetry that necessarily exists between experts making decisions concerning public health and the targets of policy. Unless it is recognized that the activities undertaken derive from human, and not merely organizational or formal, ideas and actions, there is a risk that the goal to maximize populational health overshadows the individual differences and how policy actually affect people in different ways.

However, even though it is important to keep in mind that government activities are initiated, designed, and implemented by individuals, it is equally important to acknowledge that they are also parts of collectives. Thus, both notions of individual and collective responsibility are necessary for analysis and discussion about public health risk management and communication.

In the context of societal risk management and communication, it is not a binary metaphysical question (i.e., does individual or collective responsibility exist?). Instead, both should be acknowledged and the following questions should be asked. First, to what extent is it reasonable to ascribe responsibility to individuals and the government respectively? Second, what does that responsibility involve?

In addition to detailing various aspects of risk and moral responsibility, the preceding analysis helps to frame and inform what follows. Namely, a consideration of how the coronavirus pandemic, and infectious diseases generally, can help to understand the connections between individual and collective responsibility, as well as individual and collective risk exposure and harm.

7.4 Responsibility and Public Health Risks – Examples from the Coronavirus Pandemic[2]

The coronavirus pandemic during 2020–2022[3] had severe consequences in terms of mortality and morbidity, but also economy, culture, and social activities. Organizations such as the German Ethics Council (GEC) (GEC 2020), the Malaysian Bioethics Community (MBC) (MBC 2020), and the Swedish National Council on Medical Ethics (SMER) (SMER 2020) acknowledged that the pandemic not only raises questions about the efficacy of various strategies, but difficult ethical questions as well.

Some of the most important ethical questions that arose in the context of the coronavirus pandemic concern moral responsibility. These issues are not only relevant to the coronavirus pandemic, but to pandemics and even infectious diseases generally. Echoing queries from the above sections, three important questions are: first, what is a government's primary responsibility? Second, how should both the government and individuals consider personal moral responsibility in the context of infectious diseases? Third, what is the connection between the government's responsibility and the responsibility of individuals?

7.4.1 Different Strategies – Different Values

Different countries and sometimes regions chose different strategies to respond to the pandemic. These strategies could be categorized as follows: (a) laissez-faire, (b) herd immunity, or (c) aggressive. The *laissez-faire* strategy involved few measures. The *herd immunity* strategy relied on voluntary measures. Finally, *aggressive* strategies implemented a wide range of stringent interventions, some of which entailed a limitation of civil rights (Desvars-Larrive et al. 2020, 4; Studdert and Hall 2020). For example, the governments of China, Hong Kong, Taiwan, and many countries in Europe and Africa quickly introduced aggressive strategies with strict rules regarding quarantines and lockdowns. In Chile and Argentina, policy and army forces enforced lockdowns (Thomson and Sanders 2020). Some countries relaxed the rules during the summer of 2020 but went back to more aggressive approaches in October when, once again, there was an increased spread of the infection (BBC 2021; Kuwonu 2020).

The strategies in other countries, including Brazil and initially the United States, privileged individual liberty over public health. Both Prime Minister Bolsonaro and President Trump were heavily criticized for being too laissez-faire, reacting too late, and for endorsing questionable treatments to combat the disease (Finnegan 2020).

Sweden and Norway issued voluntary, but strong, recommendations. The initial response by the UK government was similar to the Swedish strategy, introducing stricter regulations after some time had passed, for which it was criticized (Henley 2020).

The different strategies that these governments chose reflect their different views on responsibility. All strategies involve tasks that have to be implemented by the governments themselves, but the stricter the chosen strategy, the greater the task responsibility is. For example, if the government chooses to implement lockdown, mandatory face masks, vaccine passports, and so forth, the government takes it upon itself to make sure the rules are enforced and people abide by them. If the strategy is laissez-faire, the government has minimal tasks to perform. However, a herd immunity or mid-level strategy in terms of enforcement may also be seen as ascribing responsibility to individuals as well. If measures and recommendations are there, communication between governments and individuals is necessary in order for people to take on their responsibility. This is so, even if the measures are voluntary. It could be the case that a government recommending strong, but voluntary, measures is ready to make the rules stricter if people do not take their responsibility.

Management and communication during the coronavirus pandemic illustrate the importance of acknowledging the intertwinement of individual and collective responsibility. The questions discussed in previous sections become important in this context: to what extent is risk management an individual versus a collective responsibility and what does that involve? I will discuss these questions in the following sections.

7.4.2 Responsibility of Governments: Balancing Individual Rights and the Collective Good

Much like climate change and antibiotic resistance, problems related to the pandemic are not likely to be solved or managed unless both (a) states and (b) individuals take actions. The distribution of responsibility between governments and individuals is one of the main issues discussed in relation to climate change ethics and antibiotic resistance. State action is crucial, but unless there are also behavioral changes among the individuals making up the population at large, long-term change is unlikely to occur. This is vital for the current pandemic but will be even more important in order to prevent similar challenges in the future.

Regulations are important, but inadequate. Studies show that rules are not enough to change people's behavior but change of social norms and habits, etc., is also necessary. For example, lifestyle changes that are needed to reduce obesity are unlikely to succeed unless sustainable developments of habits and norms are encouraged (Fisher and Fisher 1992). These changes have to not merely be initiated but maintained. As described by Maio et al. (2007), research distinguishes between downstream and upstream interventions. Downstream interventions include, e.g., information campaigns. Upstream interventions focus on the environment and long-term change of social norms in order for desired habits to flourish, shaping conditions that "promote and sustain desired habits" (Maio et al. 2007; Verplanken and Wood 2006).

When it comes to public health problems, such as the coronavirus pandemic, responsibility for its management and communication should be shared as well. The main reason is that, unless people and governments cooperate in reaching the common goal of minimizing death and suffering, this is less likely to be achieved. The government can regulate, recommend, and create incentives and punishments, but it cannot make each and every individual abide by the rules. This is more or less true for all governments, but even more so for democratic states that need to uphold some sense of personal liberty in order to maintain legitimacy. Shared responsibility can also be justified by considerations relating to fairness. Goals like the achievement of public health and the mitigation of and adaptation to climate change are goals for populations, and populations are collectives. Arguably, governments are there to solve and manage collective problems as it is part of their rationale. The arguments for the state provided by liberal thinkers like John Locke and Thomas Hobbes are strong. While collective problems need government action, it is important to keep in mind that populations are not merely collectives but consist of individuals. Individuals have rights; in democratic states, these rights are legally codified and explicit. However, individuals also have obligations and responsibilities. Although they are allowed to make decisions concerning their own lives, they are also obligated to, for example, avoid harming others or contribute to the collective good by paying taxes. They have a right to decide about their own health, but not threaten the health of others. Furthermore, because of the fact that all states have limited resources and have to prioritize between different areas and societal problems, it is fair to expect individuals to contribute to facilitate solutions.

Thus, there are both utilitarian and deontological reasons to view responsibility for public health as shared between governments and individuals. This obviously raises questions concerning the distribution of responsibility and how to conceive the differences and boundaries between collective and individual responsibility.

When it comes to government intervention, the concepts of task and role responsibility appear reasonable and applicable. Political decisions and implementation require distribution of tasks and roles. For example, decisions have to be made concerning the role of political leaders in relation to civil servants and government agencies. These boundaries are regulated, but in terms of crisis the exact boundaries may not be as straightforward. This can result in debate and critique. This happened in Sweden, when the so-called Corona Commission criticized the government for lack of initiative at the beginning of the pandemic and stated that the government delegated too much responsibility to the Public Health Agency of Sweden (Swedish Corona Commission 2022).

Even though there may be some confusion regarding boundaries, roles and tasks are usually regulated and codified by law. However, from an ethical perspective, it is important to analyze what the roles and tasks of governments should cover. It became clear during the 2020–2022 global crisis that the pandemic entailed ethical as well as scientific and political questions.

Experts primarily engaged in communication about the coronavirus pandemic in numerical and scientific terms, for example, describing the importance of "flattening the curve." This is consistent with research showing that experts consider risk to be acceptable if the benefits outweigh the harms, and risk is calculated at the population level. But, as described before, laypeople tend to take other values into consideration, such as whether a risk was voluntary and whether risks and benefits are distributed fairly. As philosophers have argued, at least some of these values are not only important for effective communication, but for good normative reasons. Against this background, it is important that communicators take these ethical values into account. Doing that requires a balancing act between several important values, for example, reducing mortality while upholding individual autonomy. It is important that the balancing act is made explicit. This is vital in order to achieve effective communication, but also for reasons relating to legitimacy, trust, and accountability. As an example, there were instances in which experts and leaders stated that the reason for, e.g., not prohibiting all cultural events, was because there was not enough evidence of its effectiveness (Árvai and Rivers 2014; Jakobsson 2020; Nihlén Fahlquist 2018). Yet, such cases could instead be described as partly a matter of what appears to be effective and partly about the value of quality of life and the right to maintain a certain measure of autonomy.

As mentioned in a report by the GEC, democratic legitimacy requires that public health policy not be delegated exclusively to scientists, but rather should consider values. Given the complexity of making public policy decisions concerning the pandemic, "a showdown between public health imperatives and civil liberties appears inevitable" (GEC 2020).

Governments therefore have to walk a fine line between protecting the collective good and upholding ethical values. This balance is crucial due to the power imbalance between authorities and laypeople (Nihlén Fahlquist 2019). As the GEC states, governments should make sure that the fundamental rights of individuals are upheld even in cases where a utilitarian approach to "save as many lives as possible" is also a duty. It is the "democratic responsibility" of society as a whole to decide how to respond to scientific findings instead of delegating policy decisions to scientists. The core ethical issue requires taking the right measures to "sustainably safeguard a high-quality and effective healthcare system whilst, at the same time, averting or mitigating the serious adverse consequences of these measures for people and society" (GEC 2020). The MBC also identifies "the tensions between public health interests and personal rights and freedom" as a key ethical issue (MBC 2020).

The main values at stake can be evaluated using the four principles of biomedical ethics: autonomy, non-maleficence, beneficence, and justice (Beauchamp and Childress 1994). Although they are neither all-encompassing nor above criticism, these principles are useful for this analysis. For example, these principles help us see that the collective good relates more to beneficence and justice. The values often espoused by laypeople involve questions relating to autonomy and non-maleficence.

7.4.2.1 Beneficence and Justice

The principle of beneficence can be described as "a statement of moral obligation to act for the benefit of others" (Beauchamp and Childress 1994). As public health focuses on the good of the population, this principle could be seen as requiring the government to act for the benefit of the population instead of focusing on the benefit to specific individuals. Protecting and promoting the health of the population is the rationale of public health, which means that the principle of beneficence is a dominant principle. Managing the coronavirus pandemic requires the protection of as many people as possible from infection, which has the added benefit of easing pressure on healthcare systems. The principle of beneficence applies in the context of the coronavirus to the prevention of harm. Simultaneously preventing harm to vulnerable populations, the population at large, and individuals is a challenging task.

This principle of justice involves "fair, equitable, and appropriate treatment in light of what is due or owed to persons" (Beauchamp and Childress 1994). This encompasses the notions that (a) everyone should be treated equally and that (b) resources should be distributed fairly. This means that everyone should have an equal chance at good health and a long life. This relates to prevention of harm because a core problem of justice involves how to make decisions balancing the needs of patients in urgent need of healthcare and people who suffer from, for example, postponed treatments.

Some people will be harmed as a consequence of protecting others. For example, the resources needed to address the pandemic led to medical procedures being canceled or delayed. In Sweden, one of the major hospitals canceled all non-urgent surgeries due to a coronavirus outbreak (Törnquist 2020). Additional unintended side effects are increased risk of partner abuse and mental health issues (CDC 2022b).

One very important problem relates to global justice. It became clear during the coronavirus pandemic that poor countries did not have the same access to vaccines.

Clearly, the pandemic and public health raise questions both concerning global and local justice.

7.4.2.2 Autonomy and Non-Maleficence

Individual rights entail questions relating to the values of autonomy and non-maleficence. Autonomy is considered closely related to freedom of choice. Beauchamp and Childress state that respect for autonomy means acknowledging the rights of individuals to make choices and to "take actions based on their personal values and beliefs" (Beauchamp and Childress 1994). A more nuanced approach considers autonomy on a spectrum between "shallow autonomy" and "deep autonomy". For example, seatbelt requirements and stop signs would be considered infringements on autonomy if they are understood merely in relation to the freedom to act, i.e., "shallow autonomy". In contrast, the value of "deep autonomy" could be conceptualized as the value of making one's own important life choices, assessed over a longer period of time, involving "reflection on the values by which one's life will be structured" (Sneddon 2006). It entails respect for an individual's choices, but also for their capacity for conscious reflection upon these values. From this perspective, seatbelt laws are not infringements of autonomy because individuals can choose not to abide by the law (Nys 2008; Sneddon 2006).

Individual autonomy has been challenged during the coronavirus crisis, as people have been prohibited or discouraged from going outside, visiting relatives, and attending cultural events and social gatherings. Each government has had to choose between a laissez-faire or herd immunity strategy that maintains as much autonomy as possible and an aggressive strategy that prioritizes the health of high-risk groups. Regardless of the specifics, every national strategy has had to take a position along this continuum.

The principle of non-maleficence entails an obligation to refrain from harming others (Beauchamp and Childress 1994). The connection, and potential conflict, between autonomy and the right not to be harmed was described by Mill, who argued that we are allowed to do anything we want as long as it does not infringe on anyone else's right not to be harmed

(Mill 2011). Not following the rules during the coronavirus pandemic can clearly cause harm to others as a consequence.

Laypeople generally value voluntariness and autonomy in risk-taking. However, they also value fairness in risk-benefit distributions (Slovic 2000). In light of this, infringements of the right to autonomy may be considered acceptable if the underlying reason is to protect vulnerable people from being exposed to fatal risks. Against this background, the challenge for governments is to communicate the way in which autonomy and nonmaleficence are being balanced against beneficence and justice.

7.4.3 Individual Responsibility

Decreasing the spread of the pandemic also requires that individuals take personal responsibility, and as argued above, this relates to fairness as well as efficacy. As the GEC states, each individual has a responsibility to know that one's own decisions necessarily have an impact on other people (GEC 2020).

As argued above, collective problems are long term and require fundamental changes in behavior, attitudes, and decision-making on both collective and individual levels.

A useful way of conceiving this kind of long-term change comes from virtue ethics. Whereas consequentialism and deontological ethics focus on actions and individual decisions, virtue ethics departs from the entirety and complexity of people's lives. It emphasizes the importance of human beings gradually evolving virtues, which could be facilitated by one's environment, social context, and role models. Just like Jamieson argues in the context of environmental problems, there are contexts where even utilitarians should be concerned with virtues because that will be more effective (Jamieson 2007).

Conceiving responsibility as a virtue implies requiring individuals to be capable of navigating the complexity of relevant values and principles, to be able to "respond to a plurality of normative demands" (Williams 2008). The pandemic increases the complexity of responding to normative demands, with individuals having to find new ways to organize work and parenting, and to care for the elderly without compromising their health. The crisis' longevity calls for the cultivation of certain character traits and habits, such as an increased compassion for others.

However, the expectation that individuals will take responsibility for collective problems must be connected to individual contexts (Nihlén Fahlquist 2018). Having secure, well-paid employment with the opportunity to telework makes it easier to adopt new ways of living. However, insecure employment with required in-person attendance and daily use of public transport makes it more difficult. Normative demands vary widely from person to person, which influences the ability to develop personal responsibility as virtue. Some individuals who would choose to be conscientious may simply be unable to follow the rules. That said, governments can encourage the development of responsibility as a virtue by building trust and solidarity.

7.4.4 The Connection between Governmental and Individual Responsibility

Virtues can be influenced through governmental action. Two values are particularly important for facilitating a positive relationship between the government and individuals: trust and solidarity. As we have seen, solving societal problems requires both governmental and individual responsibility. Working together requires trust and solidarity between governments and individuals.

7.4.4.1 Trust

For people to be willing to take responsibility to develop the habits necessary for managing a pandemic, they need to trust their government. Trust is relational and requires that the trusted party be worthy of trust. However, trust in public health officials has decreased in the past few years. This has led to parents refusing to have their children vaccinated. There are many reasons why people mistrust vaccination, including concerns about safety and efficacy, religious beliefs, and social norms. Authorities overstating the benefits or understating the risks of vaccination have jeopardized trust. For example, during the H1N1 mass vaccination campaign in Sweden, the government called the vaccine safe. However, the vaccine caused some teenagers to develop narcolepsy. Failure to communicate the risks of new vaccines, compared to old ones, can diminish trust between laypeople and experts (Nihlén Fahlquist 2019).

Vaccination can be seen as taking responsibility for protecting others as well as oneself. Forcing people to protect others is unlikely to encourage virtuous behavior. As Dawson et al. put it, "You can compel action, but not trust." Trustworthiness in those given responsibility requires action, and if authorities are open and honest with their information, trust is more likely to be maintained (Dawson et al. 2020). The MBC states that trust is strengthened by transparency and inclusiveness (MBC 2020).

7.4.4.2 Solidarity

Interestingly, the GEC's use of the concept of solidarity is very close to that of responsibility as a virtue. It states that "[s]olidarity means the willingness to take pro-social action on the basis of relevant common ground that demands something from the person who is prepared to show solidarity". It is based on "human compassion" and a "basic feeling of togetherness" but must also be translated into actions (GEC 2020).

SMER conceptualizes responsibility as being a part of solidarity: "It is important to support those individuals who risk being hit particularly hard by infection or countermeasures, while also emphasizing individual responsibility for the choices they make in their daily life." Voluntary recommendations (instead of mandatory rules) can contribute to trust and "a sense of joint responsibility" (SMER 2020).

Dawson and Verweij distinguish between *rational* and *constitutive* solidarity. The former is based on the idea that a threat to one individual is also a threat to everyone, which entails a focus on collective societal preparation for, and prevention of, pandemics. Rational solidarity has provided the justification for many of the measures taken during the coronavirus crisis. Constitutive solidarity describes voluntary action of people to help others, i.e., taking responsibility. Citizens in countries encouraging social distancing without enforcing it could be seen as encouraging this kind of solidarity (Dawson and Verweij 2012). It is difficult to know the exact effect of policies such as police enforcement of social distancing on people's feelings of solidarity. If the crisis can be resolved soon, countries that focus on rational solidarity and promote a sense of obligation to others may find this moderate approach to be sufficient. However, the longer a crisis like this one lasts, the greater the need for people to feel connected, which creates an even greater need for individuals to develop responsibility as a virtue.

7.5 Conclusion

As this chapter showed, the concepts of risk and responsibility are closely connected, and both concepts are complex. This chapter started by building a conceptual framework focused on elements such as public health, risk ethics, virtue, individual and collective responsibility, fairness and efficacy, as well as backward- and forward-looking responsibility. This framework provided a basis for analyzing the pandemic from the perspective of risk and responsibility, as well as the ways in which the coronavirus pandemic brings issues concerning responsibility for collective problems to the fore. Collective problems, as public health threats are, require that governments as well as individuals take action and reflect on their role in relation to the achievement of the common good. I have argued that both governments and individuals need to take responsibility for reasons relating to both efficacy and fairness. Governments must communicate clearly (a) how they balance conflicts between collective health and individual rights and values and (b) what the chosen strategy entails in terms of collective and individual responsibility. These tasks require an open public discourse about the values involved. While experts can provide numbers and facts, individuals need to be involved in determining which decisions are made and how decisions are made. Success requires attention to ethical values from all involved. Individuals need to develop new character traits to help manage this pandemic and to prevent new ones. Governments must facilitate the development of such character traits by building trust and solidarity with and among citizens.

Notes

[1] On January 30, 2020, the World Health Organization (WHO) declared the global outbreak of COVID-19 to be a public health emergency of international concern (PHEIC). On May 5, 2023, the WHO declared an end to COVID-19 as a PHEIC.

[2] This section is based on, but adapted from, a previously published article: Jessica Nihlén Fahlquist (2021) "The Moral Responsibility of Governments and Individuals in the Context of the Coronavirus Pandemic." *Scandinavian Journal of Public Health* 49 (7): 815–20. https://journals.sagepub.com/doi/full/10.1177/1403494821990250 This was published under a CC-BY Creative Common License, https://creativecommons.org/licenses/by/4.0/

[3] The pandemic is under control as of March 2022, but there is uncertainty concerning whether this is temporary or not.

References

Árvai, Joseph, and Louie Rivers (Eds.). 2014. *Effective Risk Communication*. London: Routledge.

Asveld, Lotte, and Sabine Roeser (Eds.). 2009. *The Ethics of Technological Risk*. London: Earthscan.

BBC News. 2021. "Covid: How is Europe lifting lockdown restrictions?" *BBC News*. June, 25. https://www.bbc.com/news/explainers-53640249 (accessed 31 March 2022).

Beauchamp, Tom, and James Childress. 1994. *Principles of Biomedical Ethics*. 7th ed. Oxford: Oxford University Press.

Beck, Ulrich. 1992. *Risk Society: Towards a New Modernity*. London: Sage.

Bovens, Mark. 1998. *The Quest for Responsibility. Accountability and Citizenship in Complex Organisations*. Cambridge: Cambridge University Press.

Braveman, Paula, and Laura Gottlieb. 2014. "The Social Determinants of Health: It's Time to Consider the Causes of the Causes." *Public Health Reports* 129 (1): 19–31.

Centers for Disease Control and Prevention (CDC). 2022b. Coping with Stress. https://www.cdc.gov/mentalhealth/stress-coping/cope-with-stress/index.html (accessed 30 March 2022).

Cooper, David. 1968/1991. Collective Responsibility. In *Collective Responsibility: Five Decades of Debate in Theoretical and Applied Ethics*, edited by Larry May and Stacey Hoffman, 35–46. Lanham: Rowman & Littlefield.

Dawson, Angus (Ed.). 2011. *Public Health Ethics: Key Concepts and Issues in Policy and Practice*. Cambridge: Cambridge University Press.

Dawson, Angus, Ezekiel J. Emanuel, Michael Parker, Maxwell J. Smith, and Teck Chuan Voo. 2020. "Key Ethical Concepts and Their Application to COVID-19 Research." *Public Health Ethics* 13 (2): 127–32.

Dawson, Angus, and Marcel Verweij. 2012. "Solidarity: a Moral Concept in Need of Clarification." *Public Health Ethics* 5: 1–5.

Desvars-Larrive, Amelie, Elma Dervic, Nina Haug, Thomas Niederkrotenthaler, Jiaying Chen, Anna Di Natale, Jana Lasser, et al. 2020. "A Structured Open Dataset of Government Interventions in Response to COVID-19." *Scientific Data* 7 (1): 285.

Eduardo, Thomson and Philip Sanders. 2020. "Chile charts new path with rolling lockdowns, immunity cards." *Bloomberg*. https://www.bloomberg.com/news/articles/2020-04-22/with-immunity-cards-and-rolling-lockdowns-chile-forgesown-path (accessed 30 March 2022).

Finnegan, Conor. 2020. Trump and Brazil's Bolsonaro Both Downplayed Coronavirus. Now Brazil Faces a US Travel Ban. https://abcnews.go.com/Politics/trump-brazils-bolsonaro-downplayed-coronavirus-now-brazil-faces/story?id=70883803 (accessed 30 March 2022).

Fisher, Jeffrey D., and William A. Fisher. 1992. "Changing AIDS-Risk Behavior." *Psychol Bulletin* 111 (3): 455–74.

German Ethics Council. 2020. Ethikrat.org. Available at: https://www.ethikrat.org/en/press-releases/

Giddens, Anthony. 1999, "Risk and Responsibility." *The Modern Law Review* 62 (1): 1–10.

Guttman, Nurit, and Charles T. Salmon. 2004. "Guilt, Fear, Stigma and Knowledge Gaps: Ethical Issues in Public Health Communication Interventions." *Bioethics* 18 (6): 531–52.

Hansson, Sven Ove. 2003. "Ethical Criteria of Risk Acceptance." *Erkenntnis* 59: 291–309.

Hansson, Sven Ove. 2007. "Risk and Ethics: Three Approaches." In *Risk: Philosophical Perspectives*, edited by Tim Lewens, 21–35. London: Routledge.

Hansson, Sven Ove. 2013. *The Ethics of Risk: Ethical Analysis in an Uncertain World*. Houndsmills: Palgrave Macmillan.

Henley, Jon. 2020. 'Complacent' UK Draws Global Criticism for Covid-19 Response. https://www.theguardian.com/world/2020/may/06/complacent-uk-draws-global-criticism-for-covid-19-response-boris-johnson (accessed 30 March 2022).

Hermansson, Héiène, and Sven Ove Hansson. 2007. "A Three-Party Model Tool for Ethical Risk Analysis." *Risk Management* 9 (3): 129–44.

Jakobsson, Hanna. 2020. Därför vill Folkhälsomyndigheten inte stänga skolorna. https://www.dn.se/nyheter/sverige/darfor-vill-folkhalsomyndigheten-inte-stanga-skolorna/ (accessed 30 March 2022).

Jamieson, Dale. 2007. "When Utilitarians Should Be Virtue Theorists." *Utilitas* 19: 160–83.

Kuwonu, Franck. 2020. "As COVID-19 Cases Rise, African Countries Grapple with Safely Easing Lockdowns." *Africa Renewal*. https://www.un.org/africarenewal/magazine/june-2020/coronavirus/covid-19-africa-cases-rise-along-economic-hardship-countries-grapple-safely-easing (accessed 30 March 2022).

Lewis, Hywel D. 1948/1991. "Collective Responsibility." In *Collective Responsibility: Five Decades of Debate in Theoretical and Applied Ethics*, edited by Larry May and Stacey Hoffman, 17–34. Lanham: Rowman & Littlefield.

Maio, Gregory R., Bas Verplanken, Antony S. R. Manstead, Wolfgang Stroebe, Charles Abraham, Paschal Sheeran, and Mark Conner. 2007. "Social Psychological Factors in Lifestyle Change and Their Relevance to Policy." *Social Issues and Policy Review* 1: 99–137.

Malaysian Bioethics Community. 2020. Bioethics and Covid-19: Guidance for Clinicians, 1st ed. https://www.researchgate.net/publication/341396598_Bioethics_COVID-19_Guidance_for_Clinicians_1st_Edition (accessed 19 October 2020).

Mill, John Stuart. 2011. *On Liberty*. Luton: AUK Authors.

Nihlén Fahlquist, Jessica. 2016. "Experience of Non-Breastfeeding Mothers, Norms and Ethically Responsible Risk Communication." *Nursing Ethics* 23 (2): 231–41.

Nihlén Fahlquist, Jessica. 2018. *Moral Responsibility and Risk in Modern Society – Examples from Emerging Technologies, Public Health and Environment*. London: Routledge.

Nihlén Fahlquist, Jessica. 2019. "Public Health and the Virtues of Responsibility, Compassion and Humility." *Public Health Ethics* 12 (3): 213–24.

Nihlén Fahlquist, Jessica. 2021. "The Moral Responsibility of Governments and Individuals in the Context of the Coronavirus Pandemic." *Scandinavian Journal of Public Health* 49 (7): 815–20.

Nihlén Fahlquist, Jessica, and Sabine Roeser. 2011. "Ethical Problems With Information on Infant Feeding in Developed Countries." *Public Health Ethics* 4 (2): 192–202.

Nys, Thomas RV. 2008. "Paternalism in Public Health Care." *Public Health Ethics* 1 (1): 64–72.

Roeser, Sabine. 2017. *Risk, Technology, and Moral Emotions*. London: Routledge.

Roeser, Sabine, Rafaela Hillerbrand, Per Sandin, Martin Peterson (Eds.). 2012. *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*. Dordrecht: Springer.

Royo-Bordonada, Miguel Ángel, and Begoña Román-Maestre. 2015. "Towards Public Health Ethics." *Public Health Reviews* 36 (3): 1–15.

Slovic, Paul. 2000. *The Perception of Risk*. London: Routledge.

Sneddon, Andrew. 2006. "Equality, Justice, and Paternalism: Recentreing Debate About Physician-Assisted Suicide." *Journal of Applied Philosophy* 23 (4): 387–404.

Studdert, David M., and Mark A. Hall. 2020. "Disease Control, Civil Liberties, and Mass Testing – Calibrating Restrictions During the Covid-19 Pandemic." *The New England Journal of Medicine* 383 (2): 102–4.

Swedish Corona Commission. 2022. SOU 2021:89. https://coronakommissionen.com/wp-content/uploads/2021/10/summary-sweden-in-the-pandemic.pdf (19 October 2022).

Swedish Council on Medical Ethics. 2020. Ethical Choices in a Pandemic. https://smer.se/wp-content/uploads/2020/08/smer-2020_3-english-report_webb.pdf (accessed 19 October 2020).

Törnquist, Hanna. 2020. Storökning på kort tid av smittspridning i Uppsala. https://www.svd.se/operationer-stoppas-pa-akademiska-i-uppsala (accessed 30 March 2022).

Van de Poel, Ibo, and Jessica Nihlén Fahlquist. 2012. Risk and Responsibility. In *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, edited by Sabine Roeser, Rafaela Hillerbrand, Per Sandin, and Martin Peterson, 878–907. Dordrecht: Springer.

Van de Poel, Ibo, Jessica Nihlén Fahlquist, N. Neelke Doorn, Sjoerd Zwart, and Lambèr Royakkers. 2011. "The Problem of Many Hands: Climate Change as an Example." *Science and Engineering Ethics* 18 (1): 49–67.

Van Hooft, Stan. 2006. *Understanding Virtue Ethics*. Chesham: Acumen.

Velasquez, Manuel. 1983/1991. "Why Corporations Are Not Responsible For Anything They Do." In *Collective Responsibility: Five Decades of Debate in Theoretical and Applied Ethics*, edited by Larry May and Stacey Hoffman, 111–32. Lanham: Rowman & Littlefield.

Verplanken, Bas, and Wendy Wood. 2006. "Interventions to Break and Create Consumer Habits." *Journal of Public Policy & Marketing* 25 (1): 90–103.

Wikler, Daniel. 2002. "Personal and Social Responsibility for Health." *Ethics and International Affairs* 16 (2): 47–55.

Williams, Glanville. 2008. "Responsibility as a Virtue." *Ethical Theory and Moral Practice* 11 (4): 455–470.

World Health Organization. 2022. WHO Remains Firmly Committed to the Principles Set Out on the Preamble to the Constitution. https://www.who.int/about/governance/constitution (accessed 30 March 2022).

8 Responsible Risking, Forethought, and the Case of Human Gene Editing

Madeleine Hayenhjelm

8.1 Introduction

Is there such a thing as "responsible" risking? Risky decisions are often colloquially criticized for being "irresponsible" and various morally attractive approaches to risks could plausibly be described as "responsible." A recent example of this normative use of "responsible" can be found in the debate on the controversial issue of human germline gene editing: the possibility to make heritable changes to human DNA in embryos made possible by CRISPR.[1] For example, the high-profile names of the organizing committee of the gene editing summit made a statement in 2015 that it would be "irresponsible to proceed" with germline gene editing unless safety issues had been resolved and public consensus had been achieved (National Academies of Sciences, Engineering, and Medicine 2015). When Jiankui He in 2018 announced that he had edited the genome of two twin girls who had just been born, this was quickly condemned by a unified scientific community and criticized for being "irresponsible" (National Academies of Sciences, Engineering, and Medicine 2018).

What is interesting about the notion of "responsible risking" is that it points toward a potential middle category of moral advice: somewhere between advice based on moral certainty and advice based on measures of mere precaution. Or, so I shall argue. It does not require moral certainty about what is the morally best thing to do objectively all things considered. It merely requires that we do what is the wise thing given what is known and what can be done. If this is right, then it could provide action-guidance before we know for sure what is and is not morally right to do, and it could also provide action-guidance when many actions are permissible but not all are "responsible." Additionally, it could provide a measure of caution that would be less permissive than standard utility maximization but more permissive than some versions of the precautionary principle. If there is something to this idea, and the idea of responsible risking is not just trivial or redundant, it could expand our toolbox for moral advice

DOI: 10.4324/9781003276029-11

in interesting ways. Given the uncertainties that come with moral uncertainty on the one hand and the uncertainties about risks on the other, this could turn out to be very useful.

Even if there is a general case for responsible actions, there may not be a case for responsible *risking*. To act in risky ways or impose risks on others is to act in ways that could go wrong, could cause harm, and could cause damages. This is what risking means. All of this seems contrary to "responsible" actions that imply some level of caution and forethought. In fact, there is a debate that argues that we have a right not to have risks imposed upon us (McCarthy 1997; see also, e.g., Hayenhjelm and Wolff 2012; Holm 2016; Steigleder 2018). Thus, there is at least room for an argument that if we violate, or at least infringe, the rights of others when imposing risks upon them, then it is also irresponsible to do so. On the other hand, if risk impositions sometimes are morally permissible, and it seems hard to avoid that conclusion if we want to avoid a problem of paralysis (Hayenhjelm and Wolff 2012), then it seems sensible that morality would ask us to impose such risks with great care, forethought, and in ways that we would generally describe as "responsible." It seems wise to only impose risks with a certain degree of constraint, to not impose risks lightly, recklessly, or carelessly. Acting "responsibly" comes with precisely such connotations: to impose a risk "responsibly" implies that we proceed on good grounds, in careful ways, and are cautious not to bring about unnecessary harm. Thus, there seem to be enough intuitive grounds for a notion of "responsible risking" such that it is worth investigating the matter further.

I shall in the following address this question: Is there a case for responsible risking as a normative fruitful concept that could provide moral guidance when it comes to risk impositions? The main claim that this chapter will defend is this. Responsible risking already entails something of a forethought condition that would require a person to think ahead and try to anticipate future events. Furthermore, there is substantive moral content in the relevant notions of responsibility that could translate into moral requirements. If we pair these two ideas together, the forethought condition and the latent moral content in requirements from responsibility, we get three moral constraints that jointly give an idea about what would and would not be the responsible thing to do when imposing risks.

The chapter is structured as follows. The first section will set the stage and say something about key terms and main assumptions for the inquiry. The second section introduces the Forethought Condition. The third section turns to different notions of responsibility relevant to our inquiry and obligations from responsibility. The fourth section translates these obligations of responsibility into moral constraints on responsible actions. The fifth section tests the three conditions on the example from human gene editing. Finally, a last section sums up the main points and conclusions.

8.2 Preliminaries

First, a note on terminology. By "risking," I mean a deliberate action that may bring about unwanted harm as one of its (reasonably direct) outcomes (cf. Hansson 2004; Oberdiek 2017). Here, the focus lies on risk impositions – that is, risking that imposes a risk of harm on others. Thus, I impose a risk on you if and only if I act in a way that introduces a new source of harm (that may or may not result as a direct consequence of my action), or I act in a way that increases a risk of harm (that may or may not result as a direct result of my action), or in other ways that contribute to or increase risk of harm in some more indirect way.[2]

"Responsibly" refers to an evaluative and prescriptive notion beyond descriptive notions of responsibility. It is here understood as a thick moral concept. To act and risk responsibly is to live up to some substantive idea about being responsible and acting in a way that morally responds to descriptive responsibility in a morally good way.

By "responsible risking," I mean risking that is performed in a manner that we would consider responsible and "responsible" is understood as a thick moral concept. In other words, we act responsibly if we act in a way that a responsible person would have acted or ought to have acted.

Next, a note on temporality. Risking is about what has not yet happened. To impose a risk is to act in a way that could cause harm in the future. This is what makes the action risky, but whether it in fact will cause harm is not known at the time of action. By contrast, liability, blame, reparations, answerability, and similar notions of responsibility refer to what has already happened as a result of someone's actions – and often in light of what was known at the time of action. When we are inquiring into a notion of responsible risking, we are looking for a way to impose risks in a morally decent way that will align with the fact that we will be responsible for the outcomes but cannot be certain about what will be.

Both risk impositions and moral responsibility move across three nodes in time: (T1) the time of decision, (T2) the enduring phase of the action set in motion and the time when complications may arise, and (T3) the time after the effects. For some actions, T1 and T2 may be almost identical, as when knocking over a vase. For other actions, such as long-term policy decisions, T2 can extend over decades or even centuries. Part of the challenge of making a sound decision at T1 is that what seems wise to do at T1 may no longer seem that way at T3. If we may make the wrong decision at T1, we may not be able to give any satisfactory explanation of our actions at T3.

The notion of "responsible risking" needs to operate across all three nodes of time. A person acts responsibly only if they at T1, as far as reasonably possible at the time, properly consider and prepare for what could arise at T2 such that they can reasonably handle relevant situations that may arise, and properly consider and prepare for what may be the result at T3 such that they can accept responsibility for those outcomes and, as far as reasonably possible, take on the obligations that follow from those outcomes.

8.3 The Forethought Condition

Lucas (1993) provides a perfectly good illustration of what this temporal component might look like in terms of responsibility as answerability:

> If I accept that I may be legitimately questioned, I shall have that possibility in mind, and consider what answers I should give to questions that may be asked of me. I shall think about what I am doing, rather than act thoughtlessly or on impulse, and act for reasons that are faceable rather than ones I should be ashamed to avow.
>
> (Lucas 1993, 11, §1.5)

There is a stroke of genius in the above quotation. Normally, we tend to think about backward-looking responsibility as something that begins after the consequences are known: Why did you do this? What can you do to fix this? Similarly, we tend to think about forward-looking responsibility, such as role responsibility, as beginning at the point when we deliberate upon our actions as choices about what to do from then on. After the consequences are known, we hold persons to account, perhaps demand reparations, explanations, apologies, etc. Before actions are embarked upon, we tend to think about what reasons, at the time of decision, would make sense to all affected as the situation is then understood. Here, these two ideas are combined: we ought now (before the action) to consider what reasons would and would not make sense in a future scenario that could result from our actions.

We could, as part of our deliberation, planning, and acting, anticipate how our current reasons may appear to those affected by our actions should things go wrong. The notion of risk and epistemic risk complicates this picture, given that our actions may come to affect different persons other than those that we had in mind at the time of deliberation and planning, and they may require a different kind of answer. Specifically, that it never occurred to us how certain minorities would be affected by our actions would not be a satisfactory answer. In general, acting on insufficient knowledge, when relevant knowledge is hard and expensive to come by, may seem like a good idea before proceeding but serve as a very poor excuse should people come to serious harm or be discriminated against.

Hansson (2007), in a paper on risk impositions, argues that any decisions about risk ought to be guided by a kind of "hypothetical retrospection" that follows strikingly parallel ideas about forethought. When contemplating what to do, we ought to put ourselves in the hypothetical shoes of our future selves looking back at our action for each possible sets of consequences. If we pair this idea of going over all the relevant outcomes in our mind and viewing them with hypothetical retrospection, we must also keep in mind that different outcomes could affect different people, and they may require different kinds of answers. The relational aspect to answerability (see, e.g., Duff 2005; Gardner 2003) thus adds an important but challenging aspect to the forethought idea.

Both Lucas and Hansson, albeit in very different ways, point to a kind of responsible action (although Hansson does not use that word) where we *now*, before the action, need to try to imagine what kind of position we would be in, in the future, if we were morally responsible for that action and this had already occurred, and let that foresight guide our actions. In other words, to act responsibly is to anticipate what one may owe to others as a consequence of one's actions before the consequences have occurred or the action has been taken. Even if we do not know what will happen but only that things could go wrong, we can always come more or less prepared for such outcomes. One does not travel to the Himalayas without preparations. However, we will not only need to come prepared for emergencies but also for the fact that we will be the ones in charge and the ones to face those that potentially come to harm. This would imply a different kind of preparation: in the form of reasons that could justify our actions and means to repair things should they go wrong.

Perhaps we could refer to this basic idea as *the forethought condition* of responsible risking. As a point of departure, it might look something like this.

The Forethought Condition: to act responsibly, one must deliberately plan and act in a way that is compatible with later being able to deliver on one's obligations from responsibility over those actions and plans.

Stated thus, the Forethought Condition seems like a simple enough idea and relatively plausible as a core idea about responsibility. However, as with any simple idea, we need to make it a bit more precise before we can assess its implications.

One source of ambiguity is that the forethought condition as stated is vague about the temporal aspect. The obligations from responsibility at T1, T2, and T3 are not the same. However, all decisions made at T1 and T2 could potentially be relevant at T3 in terms of explanations and answers. Thus, in one reading, the obligations of primary interest would be those at T3 given that the extent to which these can be satisfied will depend upon what was done and not done at T1 and T2. If we have failed to deliver on our obligations at T1 or T2, we will, most likely, be answerable for this at T3. To make the temporal aspects more explicit, we could rephrase the forethought condition in the following way and limit our investigation here to what is responsible to do at T1 in light of one's duties at T3.

The Forethought Condition, version 2: to act responsibly at T1, one must deliberately plan and act (at T1 and T2) in a way that is compatible with being able to deliver on one's obligations from responsibility over those actions and plans at T3.

Another source of ambiguity is the vague wording in terms of moral requirements: to plan and act in a way that "is compatible with" later being able to deliver on one's obligations. More worrying than vagueness, however, is that as a general condition for responsible action it may be much too weak, and as a condition for risk impositions it may be too demanding. How strong or weak it in fact is will of course depend on how "obligations from responsibility" is to be understood. This will be discussed in the next section. Here it suffices to know that there are at least three such obligations that arise from being the person in charge; those that arise from being answerable to those affected; and to those that arise from being responsible to restore harms, losses, damages, and injuries. On the one hand, we can do many reckless things while not exactly excluding the possibility of being able to deliver on our obligations. We do not exclude the possibility of calling the fire brigade even if we start the fire by intent or accident. However, we may not be able to justify our actions, and we may not be able to repair the damages. Certain actions are not compatible with being able to provide any reasonable justification later. Thus, in combination, the three conditions may get responsible actions roughly right. However, once we turn to the matter of risk impositions, even a rather minimal account that merely requires that we do not directly undermine or foreclose our future capacity to deliver on our obligations from responsibility and are able to justify and rectify our actions may quickly become very demanding.

The challenge arises with the fact that when it comes to risk impositions, we cannot count on any one outcome as being certain but need to take all reasonably relevant possibilities into account. Furthermore, we may end up responsible for a situation we did not predict. We cannot responsibly act merely on the basis of the outcomes that we expect to occur. Many of the most damning risks introduced were not thought to be dangerous at the time of their introduction. This goes for carbon emissions, DDT, tobacco, amphetamines, plastic waste, etc. It does not seem right to thereby simply assume that because they were not foreseen, they could not have been irresponsible to introduce. Still, to claim that a risk imposition is responsible only if it is compatible with being able to deliver on obligations from responsibility across *all* possible scenarios, whether foreseen or not, seems too demanding. This would imply that we could never risk bringing about something that we later realized we could never reverse or repair or explain to those affected. Given that some risk policies could be in place over a very long time and that people could be affected in very indirect and unpredictable ways this could make any radical change, even for the better, irresponsible. Thus, we must limit the number of relevant scenarios to those that are "reasonably morally and probabilistically relevant." This, again, is a vague idea and merely points out the direction (see Oberdiek 2017 for an interesting attempt to narrow down morally relevant probability).

The Forethought Condition, version 3: to act responsibly at T1, one must deliberately plan and act (at T1 and T2) in a way that is compatible with being able to deliver on one's obligations from responsibility across all reasonably morally and probabilistically relevant outcomes at T3.

If we rephrase it in this way, the core idea becomes clear. To act responsibly, we will need to consider ways in which our actions may go wrong and have preparations for this, both in how to deal with it as it happens and in terms of fixing things and being able to explain our actions and decisions afterward. Ideally, we would be able to predict all ways in which things could go wrong, prepare for those outcomes, ensure that they would not arise and, if they were to arise, ensure that they would be properly taken care of. This is, however, unrealistic. When it comes to risk impositions, the actual risks are sometimes only learned about later. This means that for all new sources of risks, we may not even know what the relevant risks are. Thus, even the best decisions at T1 may leave one unprepared for T2 and without any good explanations at T3.

However, we need not know the precise risks in order to plan for eventualities. We may not know before the clinical trial whether a new drug is in fact efficient and safe, but we can know if the research subject has given their informed consent, whether there are medical resources available to treat side-effects, and whether there is insurance to cover unwanted outcomes, etc. When it comes to being in charge of dangerous policies, experimental practices, and other kinds of gambles that could affect others badly, such plans may be essential to what is required from responsibility. This need not apply in the same way to what is required from responsible actions in the context of a normal person going about their normal everyday affairs that still pose some level of risk. Part of this is meant to be covered by the limitation to "probabilistically and morally relevant outcomes." What in fact is probabilistically and morally relevant may be determined by moral norms of expectation – precisely of the kind that would determine when an explanation is sufficient to justify a course of action. We are not required to have an ambulance on stand-by when airing and dusting books on a balcony or to have funds set aside for the eventuality of food poisoning before inviting friends over for a seafood dinner. Part of this may have to do with the fact that this is not part of what we expect from each other and thus not something we would need to explain. Most cases of everyday risks can be remedied by apologies, simple measures of repairs (in a non-financial sense), and lessons learned for next time around. They can thus be imposed in ways compatible with later obligations from responsibility.

8.4 Obligations from Responsibility

What are the obligations presumed to follow from responsibility? To answer this question, we will first take a quick look at what responsibility in a descriptive sense roughly entails and, only then, return to what this would imply for being responsible in a more evaluative sense. The assumption is that to act "responsibly" is to be responsible, in a descriptive sense, in a good way.

Responsibility in a descriptive sense can refer to a number of different things. For example, the following four questions are all answered by a different concept of responsibility. (1) Who is in charge of A (where A is some action or domain)? (2) Who is to blame for O (where O is some outcome of an action)? (3) Who will fix O or compensate for O (where O is some outcome of an action)? (4) Why did you do A (where A is some action)?

There are various names attached to each of these categories. I will opt for role responsibility to refer to the kind of responsibility that answers the first question, blameworthiness for the second, responsibility to repair for the third, and answerability for the last one. For our purposes, we will focus on 3 and 4 and to some extent on 1 but not further discuss blameworthiness here.

Role responsibility. Who is in charge? The first concept would refer to a person being responsible by taking up or holding a position of authority over some domain or by being the moral agent who decided to act in a particular way. It connects an action to a moral agent. It could be forward-looking as in "Who will be in charge of A?" Backward-looking, as in "Who was in charge of A?" Or refer to an on-going position of responsibility over something or someone, as in "Who is in charge of this?"

Blameworthiness. Who is to blame? The second concept refers to blameworthiness, the person to praise or blame for some action. This could also extend to accountability or liability. This is a distinctly backward-looking notion of responsibility. This is also the classical notion of "moral responsibility" that dominates much of the literature on responsibility.

Responsibility to repair. The question "Who will fix this?" points to reparative responsibility, which is in a sense forward-looking, but after the consequences have occurred. It points to a role of being in charge, but for reparations rather than as author of the original action.

Answerability. The question "Why did you do it?" points to someone as being answerable for their action to others affected by it. All the first three questions can be answered by pointing to a particular person: They are responsible for A, they are to be blamed for O, and they are the one to fix O. However, the fourth question is second-personal, "Why did you do A?" is aimed at the moral agent pointed out by the other questions (cf. Darwall 2006).

The concept of responsibility as answerability has been developed in recent decades by a number of writers such as Duff (e.g., 2001, 2013), Gardner (e.g., 2003, 2008), Smith (2015), Shoemaker (e.g., 2011), and others. The basic idea is this: to be responsible is to be answerable, that is to be able to or even be obliged to provide an answer for one's action (Lucas 1993, 5). Gardner departs from the same basic idea and distinguishes between two kinds of answers: justifications and excuses – both explained in terms of reasons.

> Responsibility is what it sounds like: it is a kind of ability to respond. More precisely, it is the ability to explain oneself, to give an intelligible account of oneself, to answer for oneself as a rational being. [...] As a rational agent, one only has two ways of explaining oneself. The first is to offer a justification; the second is to offer an excuse (Gardner 2008, 123).

All four of these concepts of responsibility point to various aspects of descriptive responsibility: of what it means to be responsible for something. A person can be responsible in the sense of being the legitimate target for blame and praise. It can also refer to them being the person in charge, them being the person with obligations to repair or compensate for outcomes, or them being the person obliged to justify their actions to those affected by them. The latter three meanings of responsibility are of relevance in this context.

Responsibility in an evaluative sense would then refer to taking on, accepting, or living up to such descriptive responsibilities in a good way. Normatively, acting responsibly in this evaluative sense is also something we ought to do when faced with such descriptive responsibility. To act responsibly is to accept responsibility in a way that someone who is good at being responsible would act. I shall suggest that there is more moral content to this idea than one might first suspect.

Thus, to act responsibly is to act in ways, early on, that allows one to successfully oversee a domain over an extended period. To act responsibly is to act in ways, early on, that allows one to, later, successfully repair what one has broken or harmed. To act responsibly is to act in ways, early on, that allows one to, later, have good answers for why one did what one did. To act responsibly is to take charge of what needs to be done in light of what was done and what such actions resulted in.

This "early on" clause is essential. Just as "precaution" has an element of prevention of later harm, "responsibility" has a similar preventative element that involves planning and preparation for later events and outcomes. We want to hold people to account, to hear their explanations when things go wrong. To take responsibility is to stick around when things go wrong, to admit mistakes, and to seek to explain and repair them when they occur. To act responsibly is also to take measures and make plans such that negative outcomes are to a reasonable degree foreseen and avoided.

8.5 Responsibility Conditions as Moral Constraints

Let us return to the Forethought Condition. The general idea was this: in order to act responsibly one must seek to avoid doing, at T1 and T2, what at that point in time, we have reason to believe will foreclose our ability to be in a position to deliver our obligations from responsibility at T2 and T3. The gist of it is that we cannot claim to act responsibly and at the same time undermine our ability to do what is required from us as responsible agents.

At least three of the different parameters of responsibility above give us different kinds of failures when not satisfied: role responsibility,

responsibility to repair, and answerability. We could fail in our capacity to be charge of a situation that we have a role responsibility to be in charge of. We could fail in our capacity to "fix" what it is our role to fix. We could fail in our capacity to explain and provide reasons to those affected for actions that we have performed. What is of interest here is that these failings could provide us moral limits to responsible risking. To put it in the words of the Forethought Condition: some decisions and actions (at T1 or T2) may be such that they are incompatible with the capacity to successfully deliver on one's role responsibilities, to repair things that have gone wrong (or resulted in harm or loss), or on one's responsibility to provide satisfactory answers. If this is correct, then this gives us at least three moral limits to responsible acting from the abovementioned three kinds of responsibilities: role responsibility, responsibility to repair, and answerability.

These limits could be expressed in the following three conditions:

The control condition. We ought to act in ways that allow us to deliver on role responsibility and have control over our domain of responsibility and over outcomes.

The responsibility to fix condition. We ought to act in ways that allow us to deliver on our obligations to repair things that go wrong.

The responsibility to explain condition. We ought to act in ways that allow us to provide reasonable and acceptable answers to those affected.

The first condition would require us to act in ways that are compatible with remaining in control over what is legitimately in our domain of responsibility and enable us to deliver on decisions and actions justly required from someone in that position of responsibility. This is compatible with delegating jobs and passing responsibilities onto others. What it would rule out are various ways of not being in control while in the role of such responsibility, such as being drunk when editing genes, or withdrawing from such a role without any plan or measure for such responsibility to be delegated or taken over by someone else. Additionally, it would also rule out initiating processes that could quickly expand and become incontrollable.[3]

The second condition would require us to not act in ways that would make us unable to repair, replace, or compensate for losses and damages that we bring about. Furthermore, it would positively require us to act in ways that could contribute to our ability to repair and compensate for possible harms, losses, and injuries that we may cause. In order to do this, we must anticipate in what ways and to what extent things could be harmed or lost.

The third condition would require us to not act in ways that would render us unable to sufficiently explain and justify our actions to those directly affected by them. Furthermore, it would require us to act on reasons that would make sense and "speak" to those affected. In order to deliver on this, we must anticipate whom our actions may affect and what their crucial interests, rights, and values are.

These conditions tell us, if correct, how to impose risks responsibly: to only impose risks in ways that we can maintain control over, to only impose risks that we would be able to fix or compensate for, and to only impose risks that we could justify and explain to those affected. They also give us a hint about risks that may be, categorically, off limits. Some risks are such that they could never be, or hardly ever be, reined in if control was lost. Some outcomes are such that they could never be repaired if the worst came to be. Some risk impositions are such that they could never be justified. In all such cases, the responsible thing to do may very well be to refrain from imposing such risks.

8.6 Responsibility Conditions and Human Gene Editing

Let us try out our three conditions on the controversial case of human germline gene editing. What would responsible risking look like when it comes to gene editing and human germline gene editing? If we apply the three conditions above to gene editing and risk impositions, then we would get something like the following:

The control condition. Responsible risking requires us to only impose gene editing risks in ways that allow us to remain in control over the risks within our domain of responsibility.

The responsibility to fix condition. Responsible risking requires us to only impose gene editing risks in ways that allow us to deliver on our obligations to repair things that go wrong.

The responsibility to explain condition. Responsible risking requires us to only impose gene editing risks on grounds that we can justify to others especially to those who have a right to an answer from us.

The control condition implies that we could not responsibly impose risks that exceed what we could remain in control over within out domain of responsibility. What is implied by the control condition is something like the following: we can only responsibly put into motion courses of action that will remain largely controllable, such that it will be possible to make new decisions, change direction, etc., should new challenges, new facts, and the like arise. In the context of germline gene editing, should it become a legally permitted practice regulations, permits, licenses, professional codes of conduct, medical ethical approvals, etc., would in all likelihood help to ensure controllability. The main reason why Jiankui He was widely condemned in 2018 was the fact that he went against scientific consensus, professional codes, and regulations (Krimsky 2019).[4]

There are, however, three issues that could make it difficult to edit the human germline while not losing control over the risks and future developments as a result of this.

First, the fact that germline edits affect future individuals and that these edits are heritable.[5] The effects thus lie in the future and could extend far into the future. This means that risks could appear after the original decision-makers are gone. This need not be a worry; there are many ways to extend control responsibility across generations via regulation, institutions, and reliable processes for delegation and appointments of roles of responsibility. However, whether control and the ability to deliver upon obligations extend across generations will depend on the proportionality between capacity and size of the tasks. This could change dramatically across generations and, if much larger for later generations compared to the earlier ones, it may exceed what could be considered fair to pass on to later generations. Some decisions are easy to make (before more is known) but very hard to manage at a later stage. In the case of germline gene editing, various unknowns could make role responsibility much harder for later generations than earlier ones. We could imagine an edit that seems very promising but leads to cancer in a significant number of individuals. We could also imagine a case where radical changes to human DNA would make us much more vulnerable in a future where there are rapid and dramatic changes in the natural environment.

Second, epistemically, we do not have sufficient knowledge about effects. There are two challenges here. The first challenge stems from the fact that genes can have multiple functions. One and the same gene could thus be causal to one type of cancer and at the same time prevent another type. Furthermore, many of the diseases, conditions, and vulnerabilities that we may want to edit depend on more than one gene. This means that beyond remedying cases of severe diseases that depend on a single gene, many things could go wrong and have unwanted side-effects. The second challenge stems from the fact that some side-effects may only appear later in life or in the second generation born with edits. This means that it is difficult to gain the full epistemic picture before we, so-to-speak, try it out (Guttinger 2018). Even though it is perfectly possible, in some cases, to be in control and act responsibly when exploring the unknown, there is a limit to how far into the unknown one can venture responsibly. Part of being in control is knowing what one is doing and why. This requires some basic knowledge about the relevant outcomes. Without such knowledge, it is hard to see how we can make responsible decisions or have good plans in place.

Third, the most serious objections to germline gene editing are not about its impact on individuals but on society and what it means to be a human and regard others as fellow human beings. These kinds of objections are thus concerned about the course we would embark upon. This is an often-repeated worry that, once we embark upon germline gene editing for severe genetic disease, we will push the norm forward and lead to a slippery slope where the bottom of the slope would represent something like a *Gattica* or *Brave New World*-type dystopia (Baylis 2019; Evans 2020). The fear is that slippery slopes could lead to genetically divided societies that not only sort people according to distinct genetically determined classes but add a new genetically enhanced elite with abilities that go beyond what the best among us currently can be or achieve. This would not only drastically deepen current inequalities but also make them more permanent by having them written into our DNA. Other concerns about dramatic societal impacts are about fundamental human values lost, such as human rights premised on being "born free and equal" losing their foundation (see, e.g., Fukuyama 2003). Such developments could have large-scale impacts and develop in ways that could end up being uncontrollable. It is hard to see how we could maintain control responsibility if society or human nature is too radically or too rapidly changed.[6]

Heritability, considerable epistemic gaps, slippery slopes, and large-scale social impacts all raise challenges for the control condition – at least under current levels of knowledge. The key point is that we will be responsible for how things develop and not just for how we imagined them to develop, so we ought to be able to stay on top of that and make responsible decisions if we are to impose risks responsibly in this sense.

The responsibility to fix condition would require that we act in ways such that potential harms, damages, losses, and injuries are largely reversible, reparable, replaceable, or compensable. There are different ways of "fixing" unwanted outcomes. Should something result in an unwanted outcome, we could, potentially, reverse it, such that we are back where we were before we imposed the risk, or the outcome came about. Should we not be able to reverse it, we could repair whatever has been "broken" or, failing that, replace it with an equivalent. Sometimes there is nothing that can replace or repair something. In such cases, other things may balance the loss by offering something else that is even better. To see what this could imply for gene editing, we must first assess what could go wrong, and the nature of potential losses and harms.

First, we have the technical risks: off-target risks, unwanted but on-target risks, and mosaicism. The outcome of an edit, if proven to have unwanted side-effects, could in some cases perhaps be remedied with somatic gene therapy; but, when it comes to radical alterations, this may not be possible and will need to be "remedied" via "re-edits" of the next generation of embryos. Still, what cannot be "fixed" by re-edits could possibly be compensated for if not too grave. Not editing genes could also cause risks of harm. There is thus a reverse case, where responsibility may require us to edit an embryo to relieve it from causes for future suffering – especially in a scenario when this is an accepted practice. Here again, it may be possible to remedy this by somatic gene therapy in some cases and compensate for failing to do so in others. (These cases are further complicated by the fact that the decision to have a child at all may be conditional on the possibility of editing their DNA when there is knowledge about risk for severe genetic disease.) In general, however, the harm to an individual that may result from germline gene editing seems hard to "fix".

Second, we have societal risks, such as risks for discrimination, changed social climates and norms, the undermining of cultures, civilization and the like. These could prove even harder to "fix" given that they would require big shifts in society, and, sometimes, they may prove impossible to go back on if the shifts are too far-gone and too much has already been invested in them (Mariscal and Petropanagos 2016). Some changes to history cannot be reversed. We cannot, for instance, undo the industrial revolution. Thus, if we set in motion changes as radical as that, it may not be something we could "fix" if things went horribly wrong. However, developments could be controlled so as not to lead to such radical developments.

Third, we have the existential risk, such as the potential loss of humanity as a kind (cf. Annas, Andrews, and Isasi 2002). This may seem a highly improbable outcome. Nevertheless, should we somehow bring about the end of humanity as we know it, this would be very hard to "fix". In fact, it seems likely that this would constitute what I have referred to elsewhere as a "genuine loss" of a valuable kind – i.e., a loss that could not be repaired, replaced, or compensated for (Hayenhjelm 2018; Hayenhjelm & Nordlund forthcoming). What is in dispute is whether the loss of humanity is to count as a loss or as a gain. The transhumanists tend to think that the extinction of humanity could be thought of as a gain if replaced by a better, new species: the posthuman (Bostrom 2005). However, it is doubtful that this could be *our* gain rather than a permanent loss for us (Agar 2010; Levin 2021; Porter 2017).

In short, there are many potential outcomes that we would not be able to fix. These include harm to individuals, radical changes in norms and attitudes, potential social costs including discrimination and new genetic castes, and so on. However, most of these depend upon germline editing used for enhancement purposes. If limited to prevent medical conditions alone, or only to prevent or enhance resistance to disease, many potentially irreparable risks are avoided. However, even medically motivated edits come with risks that are not reversible for the person, and making these sufficiently safe may require trials across multiple generations and imposing risks that may not be reparable or possibly not compensable if severe enough.

Many outcomes can be repaired, losses replaced, harms healed, courses retracted, or at least compensated for. Some outcomes cannot be reversed, repaired, or compensated for – the losses are too great to make it possible to ever be outweighed (Hayenhjelm 2018). In such cases, it seems that precaution would be the right kind of approach. Some actions could render us without any means to ever "fix" the outcomes. If the consequences are severe enough, it seems that in order to be responsible, in the moral sense, to merely offer an apology or accept moral responsibility as in blameworthiness, will not suffice to make the wrong right or repair what was done (Hayenhjelm 2019). Thus, it seems that precaution and responsibility to "fix" point in the same direction: whenever there is a risk of harm or loss of such magnitude that it could not possibly be fixed, then, unless solid ground warrants exception, we cannot possibly impose such risks and be responsible while doing so. The responsible thing is to refrain from imposing such risks.

The responsibility to explain condition would require that we only impose such risks that we can reasonably explain and justify to those affected by them. We can only responsibly impose risks that we could justify and explain to others and, failing that, be able to offer some kind of excuse for.

What would and what would not count as a valid excuse or explanation in terms of germline gene editing? To a large degree, this depends on the degree of risks as well as intentions and reasons. There are, of course, obvious cases that could never be excused: such as willful and deliberate edits done with the aim of harming another person or for experimentation that would not be in the person's own interest. But even well-meaning edits that turned out to be unnecessary, riskier than thought, only relatively valuable, etc. could be questioned by the person(s) so edited.

This could rule out things like "donor siblings," or any kind of germline edits that were not for the person's own good. It would also rule out unnecessary risk-taking and acting prematurely before the risks are known and prepared against. We could not responsibly impose risks on groups of individuals that we were unaware might be negatively affected. For example, that we never considered that germline gene editing might negatively affect those with disabilities or functional variations is not a good answer to them (Sufian and Garland-Thomson 2021). More than that, as the Lucas quotation suggests, we also hold those in charge responsible for the way things are done, planned, and prepared for, as well as the number of backup plans, emergency measures, and kind of skills, training, resources, etc. that go into a responsible, but risky plan. We can accept risky and novel projects, but not sloppy and ill-prepared risky projects. "I never thought of that" is just as poor an explanation as "I never thought this would affect you," at least when "you" refers to a relevant reference group at risk. Even though not all consequences may be predictable, responsible risking would require allowances for more risks than those known about and preparations for what is not known. Thus, we may not know about all kinds of dangers in a jungle before entering it, but we could prepare as well as possible for all kinds of dangers that we can imagine. "Why did you not pack a knife?" is a perfectly good reproach even if one could not list all the possible dangers that would require a knife.[7] Responsibility does not require perfect prediction, but it does require some level of reasonable preparations that go beyond what is known based on what could happen in light of similar cases and relevant knowledge.

In many cases, the demands from control, reparability, and answerability overlap. What can be fully fixed will often be under some degree of control, and what can be fully fixed can often be excused. What could not be fixed and could not be controlled is also hard to justify. Should it turn out to be the case that experimentations with human nature came to irreparably wreck our species, undermine our core values (such as seeing the "humanity" in each other, human rights, etc.), or undermine our civilization, it is hard to see how we could fix or excuse such an outcome. This seems to hold for most "genuine losses" – in most cases, there is neither a satisfactory excuse nor a fix that could make things right. We simply took risks that were too large or acted when it was epistemically premature given the risks.[8] This is most likely where the hard limit to what could responsibly be risked lies. It should be mentioned that the three conditions could also pull in different directions. For example, we could imagine a case where some risky activity is so important that we would be held to account for not pursuing it even if we could not guarantee that we would remain in control over the events that followed.

8.7 Conclusions

Is there such a thing as "responsible risking"? This chapter has explored the notion of "responsible risking" as a thick moral concept. I have argued that the notion can be given moral content that can be action-guiding and add an important tool to our moral toolbox in the context of risk impositions. To impose risks responsibly, on the view defended, is to take on responsibility in a good way. A core part of responsible action, I have argued, is some version of a Forethought Condition. Such a condition requires us to not make decisions or plans such that we cannot deliver on our *responsibility obligations*. The morally limiting features come from what must be the case in order to be able to deliver upon one's obligations from responsibility. I have looked at three such notions: *role responsibility,*

responsibility to provide reparations, and *answerability*. All three of these hold implicit limits that can be translated into normative boundaries. I have called these *the control condition, the responsibility to fix condition*, and *the responsibility to explain condition.*

This general idea of responsible risking was tried out on the controversial case of human germline gene editing. From the control condition, we can conclude that responsible germline gene editing would require us to only impose risks in ways that allow us to remain in control over the risks within our domain of responsibility. From the responsibility to fix condition, we can conclude that responsible germline gene editing requires us to only impose risks in ways that allow us to deliver on our obligations to repair things that go wrong. From the responsibility to explain condition, we can conclude that responsible germline gene editing would require that we only impose risks in ways that we can justify to others and especially to those who have a right to an answer from us.

Are these ideas about responsible risking substantive enough to be action guiding? If so, is there anything fruitful here that is not merely trivial, redundant, or covered by the standard moral answers to risk impositions? The notion of "responsible risking" defended here points toward three distinct parameters to responsibility and thus three kinds of reasons that could support decision-making about risk impositions. Responsibility is not meant to replace other moral notions but supplement them, especially when we do not have full moral answers. We can act responsibly also when we do what later turns out to be the morally wrong thing. In fact, it is the possibility that we may do what we later could have reason to regret that makes responsibility an important notion. We can act in ways that allow us to have control over risky activities, we can be prepared to repair what could later occur, and we can act in ways that we are willing and able to explain to those to whom we may come to owe an answer. This, I have argued, would be a responsible case of risking under uncertainty or incomplete moral knowledge.

Acknowledgments

The author would like to thank the Marcus and Amalia Wallenberg Foundation. This research is one of the outcomes of the Umeå University-research project "The Rise and Risks of CRISPR/Cas9: Cultural and Ethical Perspectives on the New Gene Technology" lead by Prof. Christer Nordlund, funded by the Marcus and Amalia Wallenberg Foundation. The author would also like to thank the two dedicated editors for this volume, Adriana Placani and Stearns Broadhead, for their patience, careful close readings, insightful comments, and for the initiative of putting this interesting volume together.

Notes

1 In 2015, when it was first demonstrated in a laboratory to be possibly to apply the technology to human (non-viable) embryos, this resulted in a number of leading scientists and bioethicists and others to call for moratoria or more generally a "prudent path forward" given the high risk for off-target effects (that edits result in unintended mutations elsewhere in the DNA), unwanted on-target effects (unintended mutations at the target site) and mosaicism (incomplete edits such that some cells are and some are not edited in the intended way), and the ethical issues it raises. See, e.g., Lanphier et al. (2015) and Baltimore et al. (2015).

2 What is of relevance is both direct and some indirect consequences that are proximate enough. Here I have merely used "direct" to exclude the more far-fetched consequences. This is somewhat crude but sufficient for the purposes of this chapter.

3 The underlying intuition here runs counter to that of the Doctrine of Double Effect; the key point is not what you intend but what kind of outcomes you bring about that you could oversee and rein in if need be.

4 Outside the specific topic of germline gene editing, concerns related to what we have referred to as the control condition have been raised against bio-hackers experimenting with gene editing (largely upon themselves) and gene editing paired with gene drivers essentially making malaria-carrying mosquitoes infertile with unpredictable effects on the ecosystem.

5 The heritability aspect may also make attributions of blame difficult for those born with unwanted edits.

6 Not very surprisingly the debate on CRISPR has focused on drawing moral lines to keep the development safe enough to avoid worst outcomes but not so restrictive as to not allow medical progress. See Evans (2020) for overview on the debate.

7 The lack of explanation or excuse could cut two ways: we could end up in a position where we could not "give" any reasonable explanation to others, and we could (also) end up in a position where we had no excuse that we could accept ourselves given the outcomes. At the far end, we could end up having performed an act that was largely "unforgivable" – by our own standards, or by those affected, or by the larger moral community. It is likely that some risks that we could never repair we could explain (the reasons seemed good at the time), and that some risks that we could not justify we can still reverse or fully repair.

8 For more discussion on "genuine losses" and harder cases of risks, see Hayenhjelm (2018) and Hayenhjelm and Nordlund (forthcoming).

References

Agar, Nicholas. 2010. *Humanity's End: Why We Should Reject Radical Enhancement*. Cambridge: MIT Press.

Annas, George J., Lori B. Andrews, and Rosario M. Isasi. 2002. "Protecting the Endangered Human: Toward an International Treaty Prohibiting Cloning and Inheritable Alterations." *American Journal of Law & Medicine* 28 (2–3): 151–178.

Baltimore, David, Paul Berg, Michael Botchan, Dana Carroll, R. Alta Charo, George Church, Jacob E. Corn, et al. 2015. "A Prudent Path Forward for Genomic Engineering and Germline Gene Modification." *Science* 348 (6230): 36–38.

Baylis, Francoise. 2019. *Altered Inheritance: CRISPR and the Ethics of Human Genome Editing*. Cambridge: Harvard University Press.

Bostrom, Nick. 2005. "Transhumanist Values." *Journal of Philosophical Research* 30 (Supplement): 3–14.

Darwall, Stephen. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge: Harvard University Press.

Duff, R. Antony. 2001. *Punishment, Communication, and Community*. Oxford: Oxford University Press.

Duff, R. Antony. 2005. "Who Is Responsible, for What, to Whom?" *Ohio State Journal of Criminal Law* 2: 441–61.

Duff, R. Antony. 2013. "Relational Reasons and the Criminal Law." *Oxford Studies in Legal Philosophy* 2: 175.

Evans, John. H. 2020. *The Human Gene Editing Debate*. Oxford: Oxford University Press.

Fukuyama, Francis. 2003. *Our Posthuman Future: Consequences of the Biotechnology Revolution*. London: Profile Books.

Gardner, John. 2003. "The Mark of Responsibility." *Oxford Journal of Legal Studies* 23 (2): 157–71.

Gardner, John. 2008. "Hart and Feinberg on Responsibility." In *The Legacy of H.L.A. Hart: Legal, Political and Moral Philosophy*, edited by Matthew Kramer, Claire Grant, Ben Colburn, and Antony Hatzistavrou, 121–40. Oxford: Oxford University Press.

Guttinger, Stephan. 2018. "Trust in Science: CRISPR-Cas9 and the Ban on Human Germline Editing." *Scientific Engineering Ethics* 24 (4): 1077–96.

Hansson, Sven Ove. 2007. "Hypothetical Retrospection." *Ethical Theory and Moral Practice* 10 (2): 145–57.

Hansson, Sven Ove. 2004. "Philosophical Perspectives on Risk." *Techné: Research in Philosophy and Technology* 8 (1): 10–35.

Hayenhjelm, Madeleine. 2018. "Risk Impositions, Genuine Losses, and Reparability as a Moral Constraint." *Ethical Perspectives* 25 (3): 419–46.

Hayenhjelm, Madeleine. 2019. "Compensation as Moral Repair and as Moral Justification for Risks." *Ethics, Politics & Society* 2: 33–63.

Hayenhjelm, Madeleine and Christer Nordlund. (forthcoming). *The Ethics and Risks of Human Germline Editing: A Philosophical Guide to the Arguments*. Springer V.S.

Hayenhjelm, Madeleine, and Jonathan Wolff. 2012. "The Moral Problem of Risk Impositions: A Survey of the Literature." *European Journal of Philosophy* 20 (S1): E26–E51.

Holm, Sune. 2016. "A Right Against Risk-Imposition and the Problem of Paralysis." *Ethical Theory and Moral Practice* 19 (4): 917–30.

Krimsky, Sheldon. 2019. "Ten Ways in Which He Jiankui Violated Ethics." *Nature Biotechnology* 37: 19–20.

Lanphier, Edward, Fyodor Urnov, Sarah E. Haecker, Michael Werner and Joanna Smolenski. 2015. "Don't Edit the Human Germline." *Nature* 519 (7544): 410–11.

Levin, Susan B. 2021. *Posthuman Bliss*. Oxford: Oxford University Press.

Lucas, John. 1993. *Responsibility*. Oxford: Oxford University Press.

Mariscal, Carlos, and Angel Petropanagos. 2016. "CRISPR as a Driving Force: The Model T of Biotechnology." *Monash Bioethical Review* 34 (2): 101–16.

McCarthy, David. 1997. "Rights, Explanation, and Risks." *Ethics* 107 (2): 205–25.

National Academies of Sciences, Engineering, and Medicine. 2015. "On Human Gene Editing: International Summit Statement" (13 December 2015). https://www.nationalacademies.org/news/2015/12/on-human-gene-editing-international-summit-statement

National Academies of Sciences, Engineering, and Medicine. 2018. "Statement by the Organizing Committee of the Second International Summit on Human Genome Editing" (25 November 2018). https://www.nationalacademies.org/news/2018/11/statement-by-the-organizing-committee-of-the-second-international-summit-on-human-genome-editing

Oberdiek, John. 2017. *Imposing Risk: A Normative Framework*. Oxford: Oxford University Press.

Porter, Allen. 2017. "Bioethics and Transhumanism." *The Journal of Medicine and Philosophy* 42 (3): 237–60.

Shoemaker, David. 2011. "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121 (3): 602–32.

Smith, Angela M. 2015. "Responsibility as Answerability." *Inquiry* 58 (2): 99–126.

Steigleder, Klaus. 2018. "On the Criteria of the Rightful Imposition of Otherwise Impermissible Risks." *Ethical Perspectives* 25 (3): 471–95.

Sufian, Sandy, and Rosemarie Garland-Thomson. 2021. "The Dark Side of CRISPR." *Scientific American*. https://www.scientificamerican.com/article/the-dark-side-of-crispr/ (accessed March 20, 2022).

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part IV
Technological Context

9 Emotions, Risk, and Responsibility

Emotions, Values, and Responsible Innovation of Risky Technologies

Sabine Roeser and Steffen Steinert

9.1 Introduction

Technologies such as biotechnology, energy technologies, and digital technologies are frequently highly controversial. While such technologies often contribute to people's well-being, they can also have negative side effects or risks, which can create social disruption. Think about, for example, the polarizing effects of social media or the risks of energy technologies for health and nature. These potentially negative consequences of technologies require approaches for decision-making on how to responsibly innovate risky technologies. Technology is not value-neutral; rather, design choices imply value choices. That is why approaches to risk ethics need to include ethical values in approaches to responsible decision-making about risk (Asveld and Roeser 2009; Hansson 1989, 2012, 2013; Roeser et al. 2012), and approaches to philosophy of technology have argued for a long time that we need value-sensitive design and responsible innovation (Friedman and Hendry 2019; van den Hoven, Vermaas, and van de Poel 2015). These approaches aim to ensure that value choices are made explicitly and that these value choices are based on sound ethical considerations.

This chapter will examine the contribution that emotions and values can make to responsible innovation of risky technologies.[1] The guiding idea is that emotions can play an important role in ethical decision-making about risky technologies (e.g., Roeser 2006, 2012a, 2018). The chapter will develop this idea further and expand it to approaches of responsible innovation. The focus will be on the following key stakeholders: universities, industry, policy makers, and the public. The central idea to be investigated in this chapter is that embedding emotions and values in the innovation of risky technologies can enhance the quality of deliberation and decision-making regarding technological risks, can help to overcome stalemates, and can lead to morally and socially more acceptable and responsible technological innovations.

DOI: 10.4324/9781003276029-13

9.2 Risk, Emotions, and Values

Technological developments in, for example, energy production, robotics, biotechnology, and communication technology are taking place at a rapid pace and can have a profound impact on society, by changing our ways of life in often unpredictable ways and introducing new and unprecedented risks. For instance, it was arguably difficult to predict that social media would often negatively affect the well-being of users (Bailey et al. 2020). Public debates about such technological developments are frequently emotionally charged, resulting in stalemates between proponents and opponents (Jasanoff 2012; Siegrist and Gutscher 2010). These debates and stalemates can be explained by the fact that technological developments involve scientific information that is typically uncertain (Bammer and Smithson 2008; Slovic 2000) and because the evaluation of technology and risks involves deeply personal values and interests. Furthermore, because of their impacts on society and the environment, technological developments give rise to ethical considerations (Asveld and Roeser 2009) and emotional responses (Roeser 2010a; Slovic 2010).

The field of risk ethics has argued that decision-making about risk requires ethical reflection and public deliberation (e.g., Hansson 1989, 2012; Roeser 2007, 2018; Shrader-Frechette 1991). Mainstream approaches to risk focus on quantitative information, overlooking implicit and frequently problematic value choices (Roeser et al. 2012). Such quantitative methods also typically involve consequentialist approaches such as risk-cost benefit analysis (Sunstein 2018). However, these approaches usually overlook important issues such as distributive and procedural justice, fairness, and autonomy (Asveld and Roeser 2009).

Emotions can play an important role in highlighting such ethical issues and in deliberation about risk. However, emotions are typically considered problematic in decision-making, especially in the context of risk, as they are seen to be opposed to rationality (Dual Process Theory; e.g., Kahneman 2011, also see Sunstein 2005). Even in approaches to participatory risk assessment, emotions are not explicitly included (Roeser and Pesch 2016). While some scholars argue that emotions should be included for democratic reasons (Loewenstein et al. 2001), or because they work as an "affect heuristic" (Slovic 2010), they still think that emotions need to be corrected by rational and quantitative approaches (Slovic 2000).

In contrast to such approaches, one of us has developed an alternative approach to risk and emotions (e.g., Roeser 2006, 2018). While quantitative information is necessary in order to assess *scientific* aspects of risk, this is not sufficient to assess *ethical* aspects of risk, such as fairness, equity, and autonomy. Rather, assessing these aspects requires explicit ethical reflection, which should also involve emotions (Roeser 2006).

The plea for involving emotions in ethical reflection is grounded in a theory of risk emotions that draws on psychological and philosophical emotions research that emphasizes cognitive aspects of emotions (cf. Frijda 1986; Lazarus 1991), concerning moral emotions (Nussbaum 2001; Roberts 2003; Roeser 2011), and political emotions (cf. Kingston 2011; Nussbaum 2013; Staiger et al. 2010). Rather than seeing emotions as irrational states that disturb thinking, this approach takes people's emotions as a gateway to values (Roeser and Todd 2014). Seeing emotions primarily as irrational, biased gut reactions is a too limited view of emotions. Rather, moral emotions, in particular, can point out important moral values that should be addressed in decision-making about risky technologies (cf. e.g., Roeser 2006, 2012a, 2018). Hence, emotions are a form of practical rationality and a potential source of moral wisdom (Roeser 2006, 2009, 2010a, 2011). In that sense, emotions can be seen as "gateways to values": emotions can be an epistemological route for assessing and being sensitive to values. This is the case in more personal interactions, but also concerning political issues, as well as in the context of ethical decision-making about risk. Therefore, emotions should be explicitly included in deliberation about risky technologies, as they can draw attention to ethical considerations that get overlooked by quantitative approaches to risk. Emotions such as sympathy, empathy, compassion, enthusiasm, and indignation can highlight ethical aspects of risk such as autonomy, justice, and fairness (Roeser 2006, 2007, 2010a,b). For instance, experiencing an apprehensive emotion about a technology can highlight that the technology infringes on one's own or other people's well-being.

Of course, this does not mean that emotions are always correct; emotions can be based on misunderstandings and biases and reinforce these (Steinert and Roeser 2020; Sunstein 2010). Sometimes we are mistaken about facts, and the emotion subsides once we learn the correct information. Emotions need to be critically assessed in light of scientific information and rational, logical argumentation, as well as by emotional reflection and deliberation (Roeser 2018, Chapter 6). In other words, emotions can be an object as well as a tool of critical reflection (Roeser 2010c). This approach to risk emotions offers a fruitful alternative to current academic and practical approaches to decision-making about risk that either overlook emotions and concomitant moral values or see emotions as an obstacle to reflection. The emotional deliberation approach to technological risks sees emotions as a *starting point* for moral discussion and reflection about risk (Nihlén Fahlquist and Roeser 2015; Roeser 2012b; Roeser and Pesch 2016).

Emotions can be an important gateway to ethical considerations in value-conscious technology design (Desmet and Roeser 2015; Roeser

2012c; Roeser and Steinert 2019). However, there is no research yet on how emotions can be systematically embedded in the responsible innovation of risky technologies (for an exception, see Steinert and Roeser 2020). Further research is needed on how emotions can be systematically integrated in approaches to responsible innovation in order to address important moral values underlying emotions. In the following sections, we will set out an agenda for such research.

9.3 Emotions and Responsible Innovation of Risky Technologies

Explicitly addressing emotions and integrating them into the responsible innovation of risky technologies requires efforts by all major actors: universities and companies that develop new technologies, policy makers who develop procedures for decision-making on and policies for the regulation of innovations, and the public, concerning ways to participate in decision-making. In this section, we will provide a preliminary discussion of the potential benefits and challenges of including emotions in responsible innovation of risky technologies, and we will highlight avenues for further research. We will discuss the possible role of emotions for responsible innovation of risky technologies for four key stakeholders: universities, industry, policy makers, and the public, by reflecting on potential positive contributions as well as on potential challenges of including emotions.

9.3.1 Universities

Universities, especially universities with engineering and design schools, are key institutions when it comes to developing technologies. Not only do these institutions explicitly contribute to the creation of technology, by developing new technological innovations and providing advice and skill, but they also shape new generations of engineers and designers. This means that universities can play an important role in contributing to more responsible innovation of potentially risky technologies and the shaping of future engineers into responsible innovators.

However, assuming this role requires explicit attention to values and ethical considerations in engineering research and education programs, and an overall institutional commitment to ethics. This entails is a look at how emotions and underlying value considerations can be explicitly included in engineering research and education and, in a more overarching way, at the level of university policies. This also includes investigating and assessing best practices concerning ethics in engineering research, engineering ethics teaching, and university integrity policies, for example, concerning human research ethics, and requirements for responsible innovation.

Several philosophy and ethics departments at engineering universities, especially in the Netherlands, are leaders in integrating ethics in engineering research and education, as well as integrity policies of universities of technology. However, while there is some practical experience with this, there is not yet a lot of academic research on these topics (for some exceptions cf. Koepsell et al. 2014; Van Grunsven et al. 2021). Furthermore, these approaches have not paid explicit, systematic attention to emotions as gateways to values (although see Sunderland 2014 for the treatment of the role of emotions in engineering ethics education).

Some of the challenging questions about emotions and engineering education and research are as follows: How can engineering scholars be motivated to pay attention to and include emotions and values in their research and education? How can engineering ethics education be improved by not only focusing on theoretical ethical argumentation but also on emotional considerations? How can integrity policies of universities of technology be attuned to emotional concerns in order to create ethical awareness and to bring ethical issues to the fore? These questions are especially challenging because rules, regulations, or policies are general and abstract, while attention to emotions requires context-sensitivity, and because emotions and values are often very personal. The difference between them is exactly one of the reasons why including emotions is important: because this would do justice to context-sensitive features and provide a more fine-grained understanding of the impact of technologies on people's well-being and concerns about impacts on nature.

Without policies there might be no firm commitment, especially because paying more attention to emotions and values requires breaking up the still-prevalent culture of engineering education that focuses on quantitative methods of assessment, such as cost-benefit analysis. Furthermore, policies can be necessary to change the status quo and provide guidance in cases of conflict. However, it can be hard to bridge the gap between general rules on the one hand and context-sensitive and emotional considerations on the other. This requires further research.

9.3.2 Industry

Another important key player in the development of new technologies is industry, especially high-tech companies. The paradigm approach in much of business and economics is the neoclassical approach, according to which rationality is understood as the making of self-interested choices that maximize utility. However, this view is challenged by philosophers and alternative heterodox economic theories, e.g., feminist approaches or the Austrian and Keynesian school (Chang 2014). For instance, Powell (2010) has argued that the self-interested paradigm in neoclassical economics is neither empirically nor normatively defensible.

Zooming in on companies that develop new technologies: these companies develop artifacts that impact people's life, well-being, and the choices they make. Because of this impact, tech companies would do well to take ethics more seriously. Indeed, some companies even collaborate with professional ethicists. For instance, ethics researchers in the Netherlands have worked together with private companies in collaborative research projects funded by the Dutch Research Council (NWO, for example, in a large scale funding program devoted to socially responsible innovation), and EU projects sometimes bring together ethics researchers and industrial partners as well. These projects have resulted in academic publications, as well as in more responsible and value-sensitive innovations. Furthermore, to assess the impact of their products, some major high-tech companies have installed ethics boards. However, despite such laudable initiatives, there are still important ethical challenges concerning large-scale, systematic embedding of ethics in industry. One is the problem of "ethics-washing" (Bietti 2021), where ethics is mainly for show, and the company does not actually do anything to address ethical issues. Another issue in the collaboration between industry and ethicists is that ethicists who collaborate with industry are seen with suspicion. By becoming part of the system, these ethicists allegedly do not have the distance to the organization to critically assess it anymore (see recent media coverage of the Google ethics board, which was shut down one week after formation, Lichfield and Johnson 2019). This suspicion toward ethicists could undermine the public trust in their professionalism and threaten their credibility.

Another issue is that paying attention to ethics seems to be largely a voluntary initiative. It could be argued that it is a good thing that ethics is voluntary because it then draws on the intrinsic motivation of companies. However, if the intrinsic motivation is lacking, ethical issues will not be systematically addressed. While more and more engineering universities have institutional review boards assessing research projects in terms of human research ethics (cf. Koepsell et al. 2014), this kind of assessment is not widely used in high-tech companies, even though they engage in R&D and work on projects that can have a major impact on people's well-being. Here, policymakers and regulators could step in and make ethics reviews mandatory for certain companies (more on policy makers in the next section).

Future research is needed to investigate how ethics can be systematically embedded in companies. This involves studying how ethics committees and ethics advisors could be installed or involved in the high-tech industry, without falling prey to (possibly justified) suspicion of bias. That is, how can ethics be embedded as a genuinely impactful voice rather than being overruled or absorbed by powerful forces in industry?

Furthermore, while attention to ethics already requires a big step for tech companies, paying attention to emotions will require an even more radical change of mindset, as tech companies usually pursue formal, quantitative, and supposedly rational approaches to problem-solving. This focus on quantification and rationality comes at the expense of attention to values and ethical concerns. More work is required to figure out how to integrate emotional-moral reflection in such companies as a key ingredient to decision-making. This requires novel approaches to decision-making and leadership in high-tech industries.

Not all management practices ignore values, however. Emotional-moral reflection could enhance management approaches that focus on values, such as shared value creation, which is a principle for corporate social responsibility. Proponents of the principle of shared value creation suggest that we should find ways of creating economic value that, at the same time, creates social value (Porter and Kramer 2019). Focusing on shared value creation requires that managers think of corporations as embedded in society and communities, and that they create strategies that enhance social conditions, answer societal challenges, and create value for all stakeholders. Focusing on shared value creation means moving beyond short-term economic and corporate gains and instead focusing on how to link societal and economic progress. Integrating emotions into strategies like shared value creation would bolster the success of these value-focused approaches and lead to the creation of economic value without sacrificing social and moral values.

One idea for a new way of decision-making that takes emotions and values more seriously is to give emotions more room at the workplace and in day-to-day practices. During the design and development phase of technology, designers and engineers (but also other employees involved in the process) experience emotions that can point toward neglected values. For instance, an engineer may feel uncomfortable making certain design choices to cut costs because the resulting design could be less safe for users. Giving designers and other employees an opportunity to voice their emotions and related concerns can contribute to more ethical design (cf. Roeser 2012c). This participatory process of "innovating with emotions," which takes advantage of employees' emotions that point toward values, will require some restructuring of the design process. Making these changes, however, will not only contribute to more ethical design but will also foster an open climate where employees are welcome to talk about emotions and to raise concerns, which could contribute to a more self-critical and supportive company climate.

9.3.3 Policy Makers

Policy makers play a vital role in responsible innovation of risky technologies because they develop policies to regulate these technologies and because they develop procedures for decision-making on the innovation and implementation of technologies. Addressing emotions and values in policy making in an appropriate way is challenging. Policy makers typically follow approaches that see emotions as a source of irrationality (also cf. Kahneman 2011). They either follow technocratic approaches that are based on purely quantitative information and models, thereby leaving out emotions and explicit attention to values, or they follow populist approaches that only pay attention to citizens' concerns for instrumental or populist reasons, but not as a source of substantive insight. The problem is that in those cases, there is no genuine, critical deliberation about emotions and underlying values (Roeser 2018). Alternatively, policy makers sometimes involve the public through approaches to participatory risk assessment that may also include deliberation on values. However, those approaches usually do not pay explicit attention to emotions and may thereby miss important values (Roeser and Pesch 2016). As mentioned above, in previous work, one of us has developed an emotional deliberation approach to risk to overcome this lack of attention to emotions. The emotional deliberation approach takes emotions as the starting point of moral deliberation (Roeser 2018; Roeser and Pesch 2016).

More work is needed to investigate how an approach like emotional deliberation can best be implemented in policy making. For example, some governments try to involve members of the public via citizen panels. It could be investigated how the emotional deliberation approach can be used and further developed in order to pay explicit attention to emotions as gateways to values.

Furthermore, policy makers typically use quantitative approaches to assess risks, such as CBA or QUALYs (quality-adjusted life years). However, such approaches leave out emotions and explicit consideration of values. Even though quantitative approaches to risk are intrinsically value-laden, this is typically not acknowledged and explicated, and important ethical considerations are left out of such models—for example, issues of justice and autonomy (cf. Roeser et al. 2012). One interesting avenue of exploration is how these formal (quantitative) approaches can be made more interactive, paying attention to values of different stakeholders and including ethical considerations such as capabilities, needs, justice, and fairness, not only regarding present but also future generations. For example, in the context of decision-making about the energy transition, an option could be an interactive dashboard to let members of the public deliberate about an optimal energy mix, trying out different options and seeing their implications for different people, appealing to imagination and compassion. This can also provide motivation for climate justice by making impacts of climate change more visible to people.

9.3.4 The Public

In the current literature on risk and emotion, the public is typically portrayed as emotional and is, for that reason, seen as irrational in its responses to risky technologies (e.g., Loewenstein et al. 2001; Sunstein 2005). However, as argued above, emotional responses to risky technologies should not be dismissed out of hand as irrational. Rather, emotions can be important gateways to values and should therefore play an important role in democratic decision-making about risky technologies. Including emotions is not only important for democratic and instrumental reasons, but there are also substantive reasons to include emotions in decision-making about risky technologies, as they can play an important epistemic role by shedding light on values that may otherwise be overlooked (Roeser 2006, 2018). It needs to be investigated how the inclusion of emotions in public decision-making can be fostered. Our conventional democratic tools, such as incidental voting and binary referenda, do not do justice to ethical complexities. This is why deliberative approaches to democracy emphasize the importance of deliberation and genuine exchange of viewpoints. The emotional deliberation approach to risk emphasizes the importance of emotions for this. As mentioned before, this can be combined with approaches such as citizen panels and other participatory approaches (Roeser and Pesch 2016).

A challenge is that in the age of social media, emotional responses to technologies are themselves mediated by technologies. Social media can be democratizing by providing cheap and easy access to information and communication for everyone. However, social media also has features that make it easy to manipulate emotions. For example, “trolls” can abuse platforms and the emotional reactions of other users. Furthermore, social media platforms are often designed in such a way that they stimulate certain kinds of interactions above others and reward engagement with emotional content (Steinert and Dennis 2022). In addition, the AI in the background is designed to push emotional content. These designs tend to entice poorly reflected emotions with negative ethical implications above more reflective emotions such as compassion (Marin and Roeser 2020). Last but not least, there are regular users whose goal is to mobilize crowds rather than stimulate a respectful dialogue. Hence, while online deliberation could be a way to include citizens and their emotions and values, social media may lead to skewed emotions and values. One could argue that emotional deliberation may only work offline because of its embodied nature and because social media can be too manipulative. Genuine democratic deliberation may require real encounters. However, it seems that there is nothing intrinsic to social media that would exclude it from serving as a tool for genuine deliberation. Social media could be redesigned in such a way that it fosters (emotional) deliberation. For example, social media could allow for feedback mechanisms while typing messages with possibly hurtful content (Marin and Roeser 2020). This could give users pause to think about whether they want to post something or engage in a certain discussion. Social media could be transformed into platforms for emotional-moral deliberation.

9.4 Addressing Diverging Values and Emotions in the Responsible Innovation of Risky Technologies

In the previous sections, we discussed how values and emotions could and should play an important role in responsible innovation of risky technologies in the context of different types of stakeholders—universities, industry, policy makers, and the public. However, different stakeholders (within or between generic types identified above) can hold different values and may, accordingly, have diverging emotional responses to innovations. People's values and emotions can conflict, which means that trade-offs and decisions need to be made. This also requires moral reflection on which value decisions and value trade-offs are morally justifiable. In what follows, we discuss existing approaches to how to deal with value conflicts. We will argue how such approaches can benefit from taking emotions more seriously and "emotional deliberation" in particular.

9.4.1 Value Conflicts

The design, development, and use of technology can affect a variety of different values, and people may respond to this in different ways. A value conflict within a person, an intrapersonal value conflict, occurs, for instance, when an innovation has a positive impact for one value type a person holds and a negative impact on another value type a person holds. Take electric cars as an example where an innovation can affect various values. If you strongly care about the environment, then electric cars, with their low greenhouse gas emissions, are an innovation that you will evaluate positively. In contrast, when you strongly care about your own personal resources, then the steep price of electric vehicles may bother you. In addition, the problematic social, political, and labor conditions in regions where companies harvest the rare minerals needed for electric vehicles may not sit well with you when you strongly care about the well-being of others.

An innovation can also have positive and negative implications for one and the same value type. For instance, new products for a vegetarian diet may reduce meat consumption and thus have a positive effect on the climate. At the same time, the harvest and production of the ingredients of said products may not be sustainable and negatively impact the local environment. In a situation like this, there are both negative and positive implications for values related to the environment.

Furthermore, there can be interpersonal value conflicts. Complex technologies usually affect multiple stakeholders with a variety of considerations and values. For example, an innovation can have positive impact on a value that one person endorses but negative impact on the value of another person. An interpersonal value conflict often takes the form of a value conflict between groups or communities. For instance, engineers and managers of a wind park may endorse different values than people who will live near the turbines.

Addressing and managing public value conflicts can lead to more responsible innovations and make them more legitimate because it takes stakeholder values seriously. In what follows, we will first discuss existing methods to address value conflicts. We will then explore how attention to emotions can improve these methods. We focus on value conflicts between persons, but we think our suggestions are also partly applicable to intrapersonal value conflict. In particular, our focus is on value conflicts between stakeholders of an innovation.

9.4.2 Existing Methods to Address Value Conflicts

Authors have proposed several approaches to deal with value conflicts. It is important to note that these approaches do not discuss the role of emotions. We will briefly present some existing approaches and argue that they can be improved by considering emotions.

One way to deal with a value conflict is simply to ignore it (Meijer and De Jong 2020). However, that can be morally and pragmatically problematic. People's emotions and values are then simply disregarded, thereby foregoing important ethical considerations as well as explanations for the lack of acceptance. A more constructive and morally defensible way to solve a value conflict is to change the design of the innovation and implementation strategies to include important values of stakeholders. This may also include finding novel ways of designing and implementing an innovation.

However, oftentimes it is not possible to include all values in the design and implementation of a technology. In such a case, one has to compare, rank, and trade off values and decide which values to include. Alas, making such a trade-off is not a straightforward endeavor and involves decisions about which values supersede others and which stakeholder opinions should have weight in the decision process[2]. One major problem here is how values can be compared and ranked. There are several systematic approaches for dealing with value conflicts in the design of technologies. These approaches include the well-known, but limited (see above), cost benefit analysis and so-called satisficing. In satisficing, one trades off the loss of a value with the gain in another value, but trade-offs cannot be done below a certain threshold for each value. Another way to make a value trade-off and to solve a value conflict is to re-conceptualize the values that are at stake and what design requirements are entailed to satisfy the value (van de Poel 2014). One could also deal with value conflicts with the so-called best-worst method, which assigns weights to values, thereby ranking their relative importance (van de Kaa et al. 2020).

All these approaches to addressing value conflict can benefit from paying attention to emotions, as we will discuss in the following subsection.

9.4.3 Value Conflicts and Emotions

We suggest that taking emotions into account can provide crucial access to people's values. It is our contention that emotions can play a helpful role in alleviating and potentially resolving value conflicts, in the following ways:

1. Emotions can provide crucial information as to the relative importance that people assign to values, and this information can help to make a ranking and comparison of conflicting values.
2. Taking emotions as reflections of personal values can help to focus on easily overlooked values that are implicated by the design or implementation of an innovation. By paying attention to the emotions of stakeholders, including emotions that may seem unusual, we can gain insights into underlying values that would have been overlooked otherwise. This could help to prevent interpersonal value conflicts because the underlying values can be incorporated in the design and implementation.
3. Because emotions are linked to values, an emotion conflict may be symptomatic of a deeper conflict between values. That is, when people endorse different values, they will probably have diverging emotions about an innovation. Furthermore, paying attention to emotions in the innovation process can enable people to appreciate the emotions of others and could thus help to gain insights into their values. Emotions can play a role in various ways: not only as indicators of people's personal values, but also as a "tool" to better understand the emotional responses of others. Drawing on people's compassion and sympathy can lead to a better understanding of their perspective. This can help to

prevent disagreements related to value conflicts from hitting a dead end because people talk past one another.

4. Paying close attention to emotions could also help to address and resolve value conflicts. This can be achieved by, for example, giving people the opportunity to reflect on their emotions and to assess whether their emotions genuinely reflect their values or whether the emotions are caused by some other consideration. For instance, it could be the case that the bad feeling about an innovation is caused by the management style of the company and not so much by the features of the technology.

Besides the characteristics of a technology as such (e.g., CO_2 emissions or design features like color), the way decisions are made and the way technologies are implemented (e.g., perceived procedural fairness, distribution of costs and benefits) can have implications for people's values, driving emotions, and acceptability judgments (Contzen et al. 2021). In this way, negative emotions may be caused by the proposed implementation of a technology, and these negative emotions may then spill over to what the person thinks about characteristics of the technology itself. A proper process of emotional deliberation can let people reassess their emotional responses as well as their values and let them gain understanding of different perspectives. By reflecting on and reconsidering their emotions, people may also reconsider or reinterpret the value implications of a technology and adjust their values.

Overall, incorporating emotions and paying attention to their underlying values in the design process will contribute to a socially and morally acceptable innovation because value conflicts may be prevented and resolved. Furthermore, people want to be heard and seen, and they want their values recognized. When people are given the opportunity to express their emotions and the values that underlie their emotions are taken seriously, social acceptance of technology can be facilitated.

To be clear here, uncovering the personal values and emotions that are implicated in innovations and their implementation is not sufficient. Simply put, not all considerations of values and emotions are morally justifiable; people's emotions and values can also be morally problematic. One reason is that sometimes people uphold stereotypical perceptions of others or stick to prejudices concerning technologies or the (public or private) organization that implements the innovation. This can lead people to close themselves off from new factual information or different perspectives on values. For example, grounded in some anti-government sentiment, someone may have a biased view about the administrative body tasked with implementing a technology. This could translate into an aversion regarding the technology itself. Cases like this, however, are no reason to dismiss emotions. On the contrary, by open-mindedly engaging with emotions and underlying values and incorporating them into deliberation about innovation, the influence of biased views can be revealed and may then ultimately be reduced by inviting people to also open themselves to other perspectives. This can help to avoid potential and resolve already existing value conflicts.

9.5 Conclusion

In this chapter, we have provided an overview of why and how to include emotions in the responsible innovation of risky technologies. However, it is acknowledged that emotions can be biased and problematic. Specifically, risks and challenges related to emotions can arise in the context of forecasting one's own emotions, mixed emotions, emotional recalcitrance, and collective emotions (Steinert and Roeser 2020). In other words, emotions can be appropriate but also inappropriate, and it is important to develop insights in order to evaluate and distinguish these in the context of responsible innovation of risky technologies. This requires research to identify potential pitfalls of including emotional considerations and values of important stakeholders in the responsible innovation of risky technologies. Major ethical challenges include how to take citizen's concerns into account; how to handle the powerful interests of industry and government versus those of citizens; how to embed emotions and values in democratic decision-making about the responsible innovation of risky technologies in times of social media; how to respect and maintain individual rights and genuine moral perspectives in a context of big data, sentiment analysis, and manipulation of opinions via troll farms; how to do justice to the concerns of different stakeholders concerning well-being versus sustainability in a context of climate change; how to evaluate possible diverging emotions and values of different actors and stakeholders; and how to address possibly biased emotions. These and other related challenges require further research.

Explicitly including emotions can contribute to ethical deliberation about and responsible innovation of risky technologies by highlighting important values. As discussed in this chapter, this requires further research, developing approaches for including emotions, as well as addressing potential challenges. This future research requires an iterative process between profound theoretical analysis and real-life applications and impacts. It is a promising new avenue for bringing research on risk and responsibility further.

Acknowledgments

Steffen Steinert's research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement number 788321.

Notes

[1] Although there is no agreed-upon definition, by responsible innovation we mean approaches in research and innovation that aim to avoid negative societal impact and tackle crucial societal problems. Approaches of responsible innovation systematically consider moral and social values by paying attention to, and interacting with, stakeholders that are affected by the development and embedding of technology. For more on responsible innovation, see von Schomberg (2013) and Stilgoe, Owen, and Macnaghten (2013).

[2] Because of the difficulties involved in the prioritization of values and how to make value-trade-offs, approaches seeking to address value conflicts should be supplemented with ethical theory and normative reflection (Manders-Huits 2011).

References

Asveld, Lotte, and Sabine Roeser (Eds.). 2009. *The Ethics of Technological Risk*. London: Routledge.

Bailey, Erica R., Sandra C. Matz, Wu Youyou, and Sheena S. Iyengar. 2020. "Authentic Self-Expression on Social Media Is Associated with Greater Subjective Well-Being." *Nature Communications* 11 (1): 4889.

Bammer, Gabriele, and Michael Smithson (Eds.). 2008. *Uncertainty and Risk: Multidisciplinary Perspectives*. New York: Routledge.

Bietti, Elettra. 2021. "From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics." *Journal of Social Computing* 2 (3): 266–83.

Chang, Ha-Joon. 2014. *Economics: The User's Guide*. New York: Bloomsbury Press.

Contzen, Nadja, Annika V. Handreke, Goda Perlaviciute, and Linda Steg. 2021. "Emotions towards a Mandatory Adoption of Renewable Energy Innovations: The Role of Psychological Reactance and Egoistic and Biospheric Values." *Energy Research & Social Science* 80 (October): 102232.

Desmet, Pieter, and Sabine Roeser. 2015. "Design and Emotion." In *Handbook of Ethics, Values, and Technological Design*, edited by Jeroen van den Hoven, Pieter Vermaas and Ibo van de Poel, 203–19. Dordrecht: Springer.

Friedman, Batya, and David Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge: MIT Press.

Frijda, Nico H. 1986. *The Emotions*. Cambridge: Cambridge University Press.

Hansson, Sven Ove. 1989. "Dimensions of Risk." *Risk Analysis* 9: 107–12.

Hansson, Sven Ove. 2012. "A Panorama of the Philosophy of Risk." In *Handbook of Risk Theory*, edited by Sabine Roeser, Rafaela Hillerbrand, Martin Peterson and Per Sandin, 27–54. Dordrecht: Springer.

Hansson, Sven Ove. 2013. *The Ethics of Risk: Ethical Analysis in an Uncertain World*. London: Palgrave McMillan.

Jasanoff, Sheila. 2012. *Science and Public Reason*. London: Earthscan.

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Kingston, Rebecca. 2011. *Public Passion: Rethinking the Grounds for Political Justice*. Montreal: McGill-Queen's University Press.

Koepsell, David, Willem-Paul Brinkman, and Sylvia Pont. 2014. "Human Research Ethics Committees at Technical Universities." *Journal of Empirical Research in Human Research Ethics* 9 (3): 67–73.

Lazarus, Richard S. 1991. *Emotion and Adaptation*. New York: Oxford University Press.

Lichfield, Gideon, and Bobbie Johnson. 2019. "Hey Google, Sorry You Lost Your Ethics Council, So We Made One for You." *MIT Technology Review*, April 6, 2019. https://www.technologyreview.com/2019/04/06/65905/google-cancels-ateac-ai-ethics-council-what-next/.

Loewenstein, George F., Elke Weber, Christopher K. Hsee, and Ned Welch. 2001. "Risk as Feelings." *Psychological Bulletin* 127 (2): 267–86.

Manders-Huits, Noëmi. 2011. "What Values in Design? The Challenge of Incorporating Moral Values into Design." *Science and Engineering Ethics* 17 (2): 271–87.

Marin, Lavinia, and Sabine Roeser. 2020. "Emotions and Digital Well-Being: The Rationalistic Bias of Social Media Design in Online Deliberations." In *Ethics of Digital Well-Being: A Multidiscplinary Approach*, edited by Christopher Burr and Luciano Floridi, 139–50. Dordrecht: Springer.

Meijer, Albert, and Jorrit De Jong. 2020. "Managing Value Conflicts in Public Innovation: Ostrich, Chameleon, and Dolphin Strategies." *International Journal of Public Administration* 43 (11): 977–88.

Nihlén Fahlquist, Jessica, and Sabine Roeser. 2015. "Nuclear Energy, Responsible Risk Communication and Moral Emotions: A Three Level Framework." *Journal of Risk Research* 18 (3): 333–46.

Nussbaum, Martha. 2001. *Upheavals of Thought*. Cambridge: Cambridge University Press.

Nussbaum, Martha. 2013. *Political Emotions: Why Love Matters for Justice*. Cambridge: Harvard University Press.

Porter, Michael E., and Mark R. Kramer. 2019. "Creating Shared Value: How to Reinvent Capitalism—And Unleash a Wave of Innovation and Growth." In *Managing Sustainable Business*, edited by Gilbert G. Lenssen and N. Craig Smith, 323–46. Dordrecht: Springer Netherlands.

Powell, Jeffrey. 2010. *The Limits of Economic Self-Interest*. PhD thesis. Tilburg University.

Roberts, Robert. C. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge: Cambridge University Press.

Roeser, Sabine. 2006. "The Role of Emotions in Judging the Moral Acceptability of Risks." *Safety Science* 44: 689–700.

Roeser, Sabine. 2007. "Ethical Intuitions about Risks." *Safety Science Monitor* 11: 1–30.

Roeser, Sabine. 2009. "The Relation between Cognition and Affect in Moral Judgments about Risk." In *The Ethics of Technological Risk*, edited by Lotte Asveld and Sabine Roeser, 182–201. London: Earthscan.

Roeser, Sabine (Ed.). 2010a. *Emotions and Risky Technologies*. Dordrecht: Springer.

Roeser, Sabine. 2010b. "Intuitions, Emotions and Gut Feelings in Decisions about Risks: Towards a Different Interpretation of "Neuroethics." *The Journal of Risk Research* 13: 175–90.

Roeser, Sabine. 2010c. “Emotional Reflection about Risks.” In *Emotions and Risky Technologies*, edited by Sabine Roeser, 231–44. Dordrecht: Springer.

Roeser, Sabine. 2011. *Moral Emotions and Intuitions*. Basingstoke: Palgrave Macmillan.

Roeser, Sabine. 2012a. “Moral Emotions as Guide to Acceptable Risk.” In *Handbook of Risk Theory*, edited by Sabine Roeser, Rafaela Hillerbrand, Martin Peterson and Per Sandin, 819–32. Dordrecht: Springer.

Roeser, Sabine. 2012b. “Risk Communication, Public Engagement, and Climate Change: A Role for Emotions.” *Risk Analysis* 32: 1033–40.

Roeser, Sabine. 2012c. “Emotional Engineers: Toward Morally Responsible Engineering.” *Science and Engineering Ethics* 18 (1): 103–15.

Roeser, Sabine. 2018. *Risk, Technology, and Moral Emotions*. New York: Routledge.

Roeser, Sabine, and Steffen Steinert. 2019. “Passion for the Art of Morally Responsible Technology Development.” *Royal Institute of Philosophy Supplement* 85 (July): 87–109.

Roeser, Sabine, Rafaela Hillerbrand, Martin Peterson, and Per Sandin. 2012. *Handbook of Risk Theory*. Dordrecht: Springer.

Roeser, Sabine, and Udo Pesch. 2016. “An Emotional Deliberation Approach to Risk.” *Science, Technology and Human Values* 41: 274–97.

Roeser, Sabine, and Cain Todd (Eds.). 2014. *Emotion and Value*. Oxford: Oxford University Press.

Schomberg, René von. 2013. “A Vision of Responsible Research and Innovation.” In *Responsible Innovation*, edited by Richard Owen, John Bessant and Maggy Heintz, 51–74. Chichester: John Wiley & Sons.

Shrader-Frechette, Kristin. 1991. *Risk and Rationality: Philosophical Foundations for Populist Reforms*. Berkeley: University of California Press.

Siegrist, Michael, and Heinz Gutscher (Eds.). 2010. *Trust in Risk Management: Uncertainty and Scepticism in the Public Mind*. New York: Routledge.

Slovic, Paul. 2000. *The Perception of Risk*. London: Earthscan.

Slovic, Paul. 2010. *The Feeling of Risk*. London: Earthscan.

Staiger, Janet, Ann Cvetkovich, and Ann Reynolds (Eds.). 2010. *Political Emotions*. New York: Routledge.

Steinert, Steffen, and Matthew J. Dennis. 2022. “Emotions and Digital Well-Being: On Social Media’s Emotional Affordances.” *Philosophy & Technology* 35 (2): 36.

Steinert, Steffen, and Sabine Roeser. 2020. “Emotions, Values and Technology: Illuminating the Blind Spots.” *Journal of Responsible Innovation* 7 (3): 298–319.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. 2013. “Developing a Framework for Responsible Innovation.” *Research Policy* 42 (9): 1568–80.

Sunderland, Mary E. 2014. “Taking Emotion Seriously: Meeting Students Where They Are.” *Science and Engineering Ethics* 20 (1): 183–95.

Sunstein, Cass R. 2005. *Laws of Fear*. Cambridge: Cambridge University Press.

Sunstein, Cass R. 2010. “Moral Heuristics and Risk.” In *Emotions and Risky Technologies*, edited by Sabine Roeser, 3–16. Dordrecht: Springer.

Sunstein, Cass. 2018. *The Cost-Benefit Revolution*. Cambridge: MIT Press.

van de Kaa, Geerten, Jafar Rezaei, Behnam Taebi, Ibo van de Poel, and Abhilash Kizhakenath. 2020. “How to Weigh Values in Value Sensitive Design: A Best Worst Method Approach for the Case of Smart Metering.” *Science and Engineering Ethics* 26: 475–94.

van de Poel, Ibo. 2014. “Conflicting Values in Design for Values.” In *Handbook of Ethics, Values, and Technological Design*, edited by Jeroen van den Hoven, Pieter E. Vermaas and Ibo van de Poel, 1–23. Dordrecht: Springer Netherlands.

van den Hoven, Jeroen, Pieter E. Vermaas, and Ibo van de Poel (Eds.). 2015. “Design for Values: An Introduction.” In *Handbook of Ethics, Values, and Technological Design*, 1–7. Dordrecht: Springer Netherlands.

Van Grunsven, Janna, Lavina Marin, Taylor Stone, Sabine Roeser, and Neelke Doorn. 2021. “How to Teach Engineering Ethics?: A Retrospective and Prospective Sketch of TU Delft’s Approach to Engineering Ethics Education.” *Advances in Engineering Education* 9 (4): 1–11.

10 Responsibility Gaps, Value Alignment, and Meaningful Human Control over Artificial Intelligence

Sven Nyholm

10.1 Introduction

In the “Google DeepMind Challenge Match” in South Korea in March 2016, the 18-time world champion of the sophisticated board game Go, Lee Sedol, faced off against AlphaGo, a computer program developed by Google DeepMind. Lee Sedol won the fourth out of five games. But all four other games were won by AlphaGo. It had previously been thought that although computers can beat human chess players, it would be much more difficult or perhaps even impossible to create an artificially intelligent computer program that could beat humans at the ancient game of Go. AlphaGo proved that to be wrong (Veliz 2019).

Notably, the computer program had been “training” in preparation for the game by playing numerous games against itself. Once it was time for the match, none of the humans who had developed the computer program were able to understand the strategies of AlphaGo. A human performed the part of moving the pieces on the board around. But this person did so on the basis of instructions from AlphaGo, without knowing exactly why the computer program wanted to make those particular moves. He was a little bit like the person in John Searle’s (1990) famous “Chinese Room” thought experiment who is simply following instructions, thereby being able to produce messages in Chinese, but without understanding a single word of those messages. Hence a fascinating question arises: who, if anybody, could take credit for the victory on the part of AlphaGo over Lee Sedol? None of the humans involved could have defeated Sedol, and none of them understood how exactly AlphaGo was approaching the game.[1]

Consider next a very different event, which happened almost exactly two years later, in March 2018. For the first time, a pedestrian – Elaine Herzberg – was hit and killed by an experimental self-driving car, which was operated by the ride-hailing service company Uber in Tempe, Arizona (Levin and Wong 2018). The artificial intelligence (AI) system in the car failed to recognize Herzberg as a human being in time. First, the image

DOI: 10.4324/9781003276029-14

recognition system in the car classified Herzberg as a road sign, then as a bike (she was walking with a bicycle), then as a person – and then it kept going back and forth between different classifications (Keeling 2020). No appropriate action was taken – the car did not break – and the human "safety driver" in the car also did not notice Herzberg in time. The result was that neither the AI in the car, nor the human safety driver, stopped the car in time, and the car ran into Herzberg. She sustained deadly injuries and died in the ambulance on the way to the hospital. Who was to blame for this? The safety driver in the car, the AI in the car, Uber, or who exactly (cf. Stilgoe 2019, 1–6)?

One important thing these two cases – and others like them – have in common is that they include outcomes that the humans involved are not able to fully predict or control or perhaps even understand. This gives rise to worries about so-called responsibility gaps (De Jong 2020; Matthias 2004; Nyholm 2018; Sparrow 2007). That is, sometimes it seems as if it would be appropriate to hold somebody responsible (e.g., to praise or blame them) for some outcome – or, more strongly, it might seem that we should identify somebody who can be held responsible – but there may be nobody who it is obviously right or justified to hold responsible. And perhaps nobody is willing or able to take responsibility either. Hence a gap in responsibility seemingly arises.

In this chapter, my most general aim is to identify and discuss four broad classes of responsibility gaps. I will explore crucial differences and relations among these four kinds of responsibility gaps and map them onto gaps already discussed in the literature on this topic. This overall topic relates to ethical issues about risks in at least two important ways. First, whenever there are risks – e.g., risks related to technologies – one of the issues that always comes up is the question of who should be held responsible if something goes wrong. Second, the possibility that there might not be anybody who can appropriately be held responsible in certain types of cases can itself be seen as a form of risk, since the likelihood that we might not be able to find somebody to hold responsible for something can itself be seen as a negative possible outcome. To get the discussion up on its feet, I will use two basic distinctions from the more general theory of moral responsibility: those between backward-looking and forward-looking responsibility and between negative and positive responsibility. Using those two distinctions to create a classification matrix, I end up with what I call (i) backward-looking negative responsibility gaps, (ii) backward-looking positive responsibility gaps, (iii) forward-looking negative responsibility gaps, and (iv) forward-looking positive responsibility gaps. As the examples above indicate, I am particularly interested in how these four kinds of responsibility gaps relate to developing AI technologies. Such technologies not only have a positive potential to improve our lives but they also create many different kinds of risks, where there might be unclarities regarding who should be held responsible. On the flipside, when AI technologies create good outcomes, it can also be unclear who deserves credit, as illustrated above by the AlphaGo example. But strictly speaking, we could also identify potential responsibility gaps of these four kinds in other domains as well.

I will focus in particular on the fourth above-mentioned type of gap: what I am calling forward-looking positive responsibility gaps, by which I mean gaps relating to who should take or accept responsibility for trying to bring about certain possible good outcomes. This is not an issue that has typically been discussed in these terms, but I will relate that idea to the discussion of the "value alignment" of AI (i.e., the idea that we should create AI systems that align with human values, interests, or goals) (Gabriel 2020; Russell 2019). My worry is that there might potentially be an important forward-looking positive responsibility gap related to the goal of AI value alignment. At any rate, I will use AI value alignment as a potential case in point of the fourth type of responsibility gap I will identify. As far as I am aware, the debate about technology-related responsibility gaps has not yet recognized the existence of this kind of responsibility gap, nor related it to the issue of AI value alignment.

When it comes to how and whether the four types of responsibility gaps might be filled, I will highlight two predictable asymmetries in how motivated people are likely to be to try to fill these responsibility gaps by taking responsibility for the outcomes at issue. As I will argue, depending on what kind of responsibility gap is in question, people are likely to be more or less willing to step forward and take responsibility. And sometimes when they are willing to do so, this might not be wholly appropriate. Lastly, I will also very briefly compare the prospects for two suggested "solutions" to responsibility gap worries in the existing literature: one I myself have put forward (Nyholm 2018, 2020) and another one defended by Filippo Santoni de Sio and Giulio Mecacci (2021).

These two solutions are similar in nature and can both be viewed as attempts to explain what it is to have what is sometimes called "meaningful human control" over AI systems (Santoni de Sio and van den Hoven 2018; Verdiesen, Santoni de Sio, and Dignum 2020). I will argue that these suggestions about how to fill technological responsibility gaps are most plausible when we conceive of them as theoretical idealizations or general schemas for how the problem can be solved in principle, but that they may not always be easy to convincingly and straightforwardly apply in practice to real-world cases. Perhaps more significantly for the main aim of this chapter, however, I will also argue that once we have clearly identified the four broad classes of responsibility gaps I will be discussing, we will see that there are additional challenges for these two conceptions of meaningful human control over AI that concern the relations among the different kinds of responsibility gaps. As I see things, then, identifying and clarifying the relations among the four types of responsibility gaps I am distinguishing among in this chapter is important in part because it raises new challenges for existing views about how to fill responsibility gaps.

10.2 Responsibility Gaps More Generally Considered and Four Different Kinds of Responsibility Gaps

Worries about responsibility gaps are not limited to the context of outcomes caused by, or events involving, AI systems. Another type of case in which similar worries arise is that of outcomes not caused by any identifiable individual person, but by groups, or perhaps by the cultures of organizations. For example, in his 2007 article "Responsibility Incorporated," Philip Pettit begins his discussion with the case of the MS Herald of Free Enterprise, a ferry that capsized off the coast of Belgium in 1987. This accident was blamed on a lax safety culture in the organization. It seemed like somebody should be held responsible. But it was unclear whether any individuals could be held responsible (Pettit 2007). Similarly, when the Challenger space shuttle from NASA had its disastrous accident killing seven astronauts in 1986, this was attributed, not to any individual person, but to the culture at NASA at the time (Goodpaster 2007). Such cases may potentially involve responsibility gaps. In general, then, the idea of a responsibility gap is neutral with respect to what exactly the problem might be that is giving rise to the gap. The problem might be caused by AI and other advanced technologies, but it does not have to be. It might also be caused by other things.

In the particular case of the MS Herald of Free Enterprise ferry accident, Pettit (2007) is of the view that the organization as a whole could be viewed as a corporate agent, which could be held responsible, even if no individual members of that organization could be singled out as the responsible parties. Others will look at a case like that and say that there is a "problem of many hands" there. Too many people are involved, each of whom bears some small amount of responsibility, but nobody (neither any individual nor any organization) might be the obvious main center of responsibility (Van de Poel, Royakkers, and Zwart 2015). I bring up these points partly to illustrate that the idea of responsibility gaps is not confined to the context of AI, and partly to bring up two ideas that we will later briefly return to, namely, responsibility for group agency and the so-called problem of many hands.

Let us now consider two commonly made distinctions from more general discussions about moral responsibility, which will serve as the basis for identifying the four types of responsibility gaps I distinguish among below. There is a distinction, first, between what is sometimes called *backward-looking responsibility*, on the one hand, and *forward-looking responsibility*, on the other hand (Nihlen Fahlquist 2017; Verdiesen, Santoni de Sio, and Dignum 2020). The former is perhaps the most familiar kind of responsibility: responsibility for something that has been done or that has happened in the past. Some questions about backward-looking responsibility concern whom to blame for something that has happened. Who, for example, should be blamed for the capsizing of the MS Herald of Free Enterprise or the explosion of NASA's Challenger space shuttle? Those are questions of backward-looking responsibility. But before anything bad happens, there can also be a question of whose responsibility it is to try to make sure that such an outcome does not come about. For example, there could be somebody whose responsibility it is to make sure that the proper safety precautions are in place to avoid a ferry sinking. This is a question of forward-looking responsibility: responsibility to make sure that things will happen in certain ways rather than in others.

The second distinction I want to bring up and make use of is between what I will call *positive* and *negative* responsibility (Danaher and Nyholm 2021). We do not only hold people responsible in negative ways, such as by blaming them or punishing them for bad outcomes, but we also hold people responsible in positive ways, such as when we praise or reward them in relation to good outcomes. If somebody does something good, for example, they can deserve to get credit for this. Moreover, it can be very nice for them to be singled out as the responsible person who did the good thing: again, they might be praised or rewarded. This is an example of what I am referring to by the expression "positive responsibility."

With these distinctions drawn, I am now in a position to identify the four types of responsibility gaps that I propose to distinguish. The first type of responsibility gap – the most commonly discussed type – is a *backward-looking negative responsibility gap*: such a gap occurs when something negative has happened, and it seems that somebody ought to be held responsible, but it is unclear who could justifiably be held responsible. The second type is a *backward-looking positive responsibility gap*: something good has happened, and it seems that somebody ought to be viewed as responsible for this, but it is unclear who, if anybody, could take credit or deserve to be viewed as responsible. The third type is what I call a *forward-looking negative responsibility gap*: there is a risk of a bad outcome, and it seems that there should be somebody who is responsible for taking precautions against the risked bad outcome, but it is unclear who should have or be assigned this responsibility. Lastly, the fourth type is a *forward-looking positive responsibility gap*: there is an opportunity

Table 10.1 Four types of responsibility gaps

Responsibility Gaps	*Backward-Looking*	*Forward-Looking*
Negative	Backward-looking negative responsibility gaps	Forward-looking negative responsibility gaps
Positive	Backward-looking positive responsibility gaps	Forward-looking positive responsibility gaps

for potentially producing some good outcome, and it seems that it would be appropriate that somebody should have the responsibility of trying to bring about the good outcome, but there is nobody who is a clear candidate for taking on this responsibility. These four types of responsibility gaps are shown in Table 10.1.

10.3 More on the Four Types of Responsibility Gaps and Their Relation to the Technology Ethics Responsibility Gaps Literature

Most recent discussions about responsibility gap worries have been about what I am here calling backward-looking, negative responsibility gaps. Two much-discussed contributions to the literature are Andreas Matthias' much-cited 2004 article that very clearly articulated the idea of potential responsibility gaps created by autonomously operating AI systems and Robert Sparrow's also very frequently cited 2007 article about "killer robots." By that phrase, Sparrow (2007) means autonomous weapons systems. Sparrow argues in his article that because it might be difficult or even impossible to predict and control what autonomous weapons systems or "killer robots" will do, and responsibility depends on the ability to predict and control events and outcomes, there will be responsibility gaps if we permit the use of these technologies in warfare. Alexander Hevelke and Julian Nida-Rümelin (2015), in turn, have argued that if a self-driving car is operating in a fully autonomous mode, then it might also be impossible to predict and fully control what the car will do, and so it would be unfair to hold people riding in that car responsible for any outcomes the car might produce.

Many discussions of responsibility gaps have introduced different variations on the theme of backward-looking negative responsibility gaps. John Danaher (2016), for example, argues that apparent crimes or serious injuries caused by AI systems or robotic technologies might create "retribution gaps": people may desire, or it may seem appropriate, that somebody should be punished, but there might be no moral agent who it would be just or right to punish. If this is correct, there would be responsibility gaps related both to blame and retributive punishment.

Daniel Tigard (2020), in turn, has related theoretical discussions about the different "faces of responsibility" from Gary Watson and David Shoemaker to the issue of responsibility gaps. Without going into too much detail, Watson and Shoemaker argue that there are distinctions to be drawn among the attributability, accountability, and answerability aspects of responsibility: who is the main agent behind some outcome, who should be held to account, and who can answer for what has happened? Tigard argues that when we think about AI-created responsibility gaps, we can envision responsibility gaps having to do with attributability, accountability, and answerability. These are usually discussed in relation to negative responsibility. Thus understood, they would all be related to versions of what I am calling the backward-looking negative responsibility gap.

Filippo Santoni de Sio and Giulio Mecacci (2021) discuss three kinds of backward-looking responsibility gaps of this negative type. The first two are similar to the ones mentioned above: culpability gaps and moral accountability gaps. The former are gaps with respect to who can justifiably be blamed, and the latter, as described by Santoni de Sio and Mecacci, overlap with what Tigard calls accountability and answerability gaps. The third kind of gap Santoni de Sio and Mecacci discuss they call public accountability gaps. These are gaps related to what public officials or other public figures can be held accountable for. As they are discussed by Santoni de Sio and Mecacci, those gaps also fit into the broader category of backward-looking negative responsibility gaps.[2]

What about backward-looking positive responsibility gaps? Danaher and Nyholm (2021) discussed an "achievement gap" in relation to the automation and robotization of work. If more and more tasks in many forms of work are taken over by AI systems or robots, this might interfere with people's opportunities to make important and praiseworthy achievements at work, thus threatening one important component of what we associate with meaningful work. For example, if the role of medical doctors becomes less about identifying what is wrong with their patients and more about communicating to patients what AI systems think is wrong with the patients, this might seem to undermine the human component and the important role we usually associate with being a medical doctor. Similarly, if some scientific discovery is made, not by a human scientist, but by an AI system recognizing some pattern using machine learning, it might seem as if somebody deserves credit for the discovery, but nobody might clearly deserve it. The discovery might not have directly required human ingenuity or effort. All the key work leading to the particular discovery might have been done by the pattern-seeking AI system, not any of the humans involved. This might create a gap where it seems that something praiseworthy has been done, but perhaps nobody clearly deserves to be praised.

Consider next forward-looking negative responsibility gaps. One example of this is something I have elsewhere called "obligation gaps" (Nyholm 2020, 34, 166–7; cf. Müller 2021). Suppose, for example, that a self-driving car is driving at high speed and that it seems appropriate that it should avoid hitting anybody. If there could be a backward-looking negative responsibility gap in a case where the self-driving car did hit somebody, then this might be taken to mean that there is also a forward-looking negative responsibility gap related to the apparent obligation to avoid having somebody be hit before this happens.

More generally, if morally sensitive tasks are outsourced to an AI system (e.g., an autonomous weapons system), there might be certain things it would be "wrong," so to speak, for the AI system to do. But what if the AI system is not a moral agent who could be held responsible for its "actions" if it does something wrong? If so, then a common way of thinking about obligation does not seem to apply to it. That common way of thinking about obligation is that we are obligated to do something if it would be appropriate to blame or punish us if we failed to act in the given way and we did not have an adequate excuse or justification for this omission (Darwall 2004; Gibbard 1991). If we have AI systems or robots tasked with performing morally sensitive actions, but they are not full moral agents because they cannot be held responsible for their behaviors, this might accordingly create forward-looking negative responsibility gaps. In this category, we can also put what Santoni de Sio and Mecacci (2021) call "active responsibility gaps," which also seem to fit with the idea I am calling forward-looking negative responsibility gaps. In other words, while three of their responsibility gaps fit into my first category of backward-looking negative responsibility gaps, their fourth kind fits into my category of forward-looking negative responsibility gaps.

What about what I am calling forward-looking positive responsibility gaps? As far as I can tell, this idea has not been discussed in these terms in the existing literature about AI and technology ethics. So, I cannot illustrate this idea by referring to the existing literature about responsibility gaps. However, I think that the discussion of the so-called value alignment problem of AI might be a case in point regarding this (e.g., Gabriel 2020; Russell 2019, 2020). And I think some common ways of thinking about duties of beneficence (i.e., duties to promote good outcomes) both in moral theory and in common-sense morality might give rise to a responsibility gap here. I explain these points in Table 10.2.

Here, first, is another version of Table 10.1, which incorporates what has been discussed in this section:

Table 10.2 Four types of responsibility gaps, expanded

Responsibility	*Backward-Looking*	*Forward-Looking*
Negative	Backward-looking negative responsibility gaps (e.g., responsibility gaps related to "killer robots" and self-driving cars) (Hevelke and Nida-Rümelin 2015; Sparrow 2004), retribution gaps (Danaher 2016); negative attributability, accountability, and answerability gaps (Tigard 2020); and what Santoni de Sio and Mecacci call culpability, moral accountability, and public accountability gaps	Forward-looking negative responsibility gaps (e.g., obligation gaps) (Nyholm 2020); gaps in "active responsibility" (Santoni de Sio and Mecacci 2021).
Positive	Backward-looking positive responsibility gaps (e.g., achievement gaps in the workplace where tasks have been outsourced to AI systems and robots) (Danaher and Nyholm 2021)	Forward-looking positive responsibility gaps (e.g., gaps related to who should make sure that future AI systems are aligned with human values, goals, and interests)

10.4 Forward-Looking Positive Responsibility Gaps and the AI Value Alignment Problem

Suppose that we create a powerful AI system that reliably tracks important human values, goals, or interests. Perhaps we might come up with a cancer diagnosis and treatment suggestion system that reliably makes accurate diagnoses and then suggests effective treatments. This system might be like the Go-playing AlphaGo computer program in the following respects. It might be hard to understand the inner workings of the system, but it might nevertheless reliably get the job done (cf. Robbins 2019). This is an example of what is often referred to as "value alignment": the desirable situation that AI systems fit with, or align with, human values, goals, or interests (Bostrom 2014; Gabriel 2020; Russell 2019). The question is who exactly is responsible for achieving this positive goal.

In particular, when people discuss AI value alignment, they are often discussing AI systems that we do not yet have, but that may come to be developed; and the issue is whether these powerful future AI systems will be aligned with human values, goals, or interests. If it is unclear who is responsible for helping to bring about such systems, but it seems like somebody should have such a responsibility, there is potentially a forward-looking positive responsibility gap there.

In both common sense and many influential moral theories, positive responsibilities (responsibilities to make sure that good outcomes are achieved) are typically treated as being much weaker or less stringent forms of responsibility than negative responsibilities to avoid harm or other forms of bad outcomes. In other words, while people are usually held responsible for not harming others and blamed if they harm others, it is less common to hold people responsible for positively striving to bring about good outcomes (Persson and Savulescu 2012, 4). Acting in ways that promote good outcomes is more commonly treated – both in moral theory and common sense – as something it is very nice and virtuous to do, but that is optional, i.e., not something we are held positively responsible for doing in the forward-looking sense of responsibility.[3] Ingmar Persson and Julian Savulescu capture this aspect of how most people think in a succinct way when they write the following in a discussion about moral attitudes that are deeply ingrained into ordinary common sense:

> We intuitively feel, for example, that we are more responsible for the harm we cause than for benefits we fail to cause and, thus, have moral duties or obligations not to harm, but not to benefit.
>
> (Persson and Savulescu 2012, 4)

If this is right, common sense does not posit a strong and clear responsibility to act so as to promote future good outcomes. But what about the sorts of moral theories that we find in moral philosophy, which typically try to improve upon ordinary common-sense moral reasoning?

Utilitarian or, more broadly, consequentialist theories of ethics stipulate that the right action is the one that brings about the most good overall (Driver 2006). It might be thought that such views automatically avoid the problem of forward-looking positive responsibility gaps. However, many critics of such theories argue that they are too demanding: we should not, critics say, be held responsible for going around making the world a better place (Williams and Smart 1973). And, perhaps even more importantly, many utilitarians and consequentialists themselves say that the goal of promoting the overall good is best achieved if we allow people to promote their own personal interests, so long as they do not harm other people. Both Jeremy Bentham (1789, chapter XVII) and John Stuart Mill (1859) defended that claim in their classical books about utilitarian ethics. Later on, many thinkers in that tradition have added that many human values involve partiality to oneself and one's near and dear (e.g., Pettit 2015; Portmore 2011). So, the overall good, they say, is best promoted if people are permitted to favor themselves and their near and dear. In other words, not even consequentialists tend to stipulate a strong responsibility to actively go around trying to make the world a better place.

For another example from ethical theory, consider also Kant's (1785) ethical theory. Just like consequentialists do, Kant also claims that we have a moral duty to promote the happiness of other people. However, Kant claims that this is a "wide imperfect duty": we should make the promotion of others' happiness into an end of ours, but we do not have to go around promoting this goal all the time. We do not, in other words, have a strong responsibility to make this into a priority according to Kant. Our priority, on a Kantian perspective, is to avoid treating people with disrespect or violating their dignity: we have a perfect duty – and, in other words, strong responsibility – to avoid this. We do not, in the same way, have a strong and perfect duty to promote the good of others, even if it should be among our more general aims to do so.

We can now make the following argument. Step 1: Common sense and many moral theories – like many forms of consequentialism and Kantian ethics – do not posit a strong positive responsibility to actively promote future good outcomes. Step 2: If that is so, and there are certain good outcomes that could potentially be achieved by future AI systems, this might be seen as something it would be great if somebody would strive for or help to bring about, but not something that anybody has a positive responsibility to actively work toward. Accordingly, the value alignment of future AI systems might be something that would be highly desirable, but a case where there is a forward-looking positive responsibility gap.[4]

10.5 Two Asymmetries Relating to These Responsibility Gaps

In the remaining sections, I will discuss the question of how and whether the various types of responsibility gaps can be filled. The perhaps most obvious practical "solution" that one might come up with is that somebody would step forward and take responsibility and, thereby, try to fill whatever responsibility gap there might be. However, even though this type of solution to responsibility gaps can solve the practical problem of finding somebody who is willing to take responsibility, it does not necessarily solve the normative problem of finding somebody who it is right or appropriate to hold responsible. Moreover, depending on what kind of gap we are focusing on, people can be expected to be more or less willing to step forward and take responsibility.

When one reflects on the above-identified four different kinds of responsibility gaps and how they are related to each other, it appears likely that there are going to be two noteworthy asymmetries relating to how willing people might be to step forward and volunteer themselves as responsible parties who could fill these gaps. The first asymmetry that is to be expected relates to the difference between backward-looking negative responsibility gaps and backward-looking positive responsibility gaps. The second asymmetry relates to the difference between backward-looking positive responsibility gaps and forward-looking positive responsibility gaps.

First asymmetry: After something negative has happened, people are typically unwilling to take responsibility for it. They usually prefer to blame other people instead. This is not surprising, since being held responsible for some bad outcome typically means that one has to bear some burden or suffer some set-back. This could range from criticism from other people and a guilty conscience, to something more severe such as fines, legal punishment, or even violence or other forms of attacks from other people (Mill 1863). In contrast, after something good has happened, people are typically highly motivated to present themselves as having been responsible and claim credit for the good outcome. Again, this is not surprising. Being viewed as responsible for something positive usually means receiving praise or rewards from other people, and a sense of pride in oneself. What this first asymmetry means in terms of filling responsibility gaps related to AI systems is that it will be harder to find willing candidates who want to take responsibility for bad outcomes caused by these AI systems than it will be to find willing candidates who want to take responsibility for good outcomes caused by AI systems. In the latter case, people are likely to step forward and claim credit for what was really primarily the outcome of the operations of the AI system and, in that way, attempt to fill responsibility gaps in a potentially undeserved way.

Second asymmetry: At the same time, it is likely that the following asymmetry will also be observed in practice when it comes to the difference between backward-looking positive responsibility gaps and forward-looking positive responsibility gaps. While people are likely to be highly motivated to put themselves forward as having been responsible for good outcomes that have already occurred, they are less likely to be equally motivated to put themselves forward as candidates who will take responsibility for making sure that certain good outcomes will be achieved. Again, the differences in “costs” associated with the two different kinds of responsibility are relevant here. It may cost you very little or nothing to take responsibility for something good that has happened, and you may gain from this: you may win people’s esteem, praise, and various rewards. In contrast, putting oneself forward as a candidate for being responsible for the promotion of some positive future outcome can involve taking on costs, including opportunity costs, since one could otherwise be devoting one’s time to other things than trying to bring about the outcomes in question, which might be demanding, difficult, and otherwise expensive.

Of course, there are some forward-looking positive responsibilities that people happily take on. For example, people are usually willing to take responsibility for making sure that things go well for their children (Persson and Savulescu 2012). But when it comes to securing other future benefits, such as benefits that will be enjoyed by strangers or benefits for future generations of people, people are usually much less willing to accept responsibility for them than they are willing to accept responsibility for good things that have happened in the past. This applies to companies as well, including tech companies of the sorts that will be producing many of the AI technologies that ought to be aligned with human values, goals, and interests. The average corporation does not typically eagerly and willingly take on positive responsibilities to make sure that their products promote good outcomes for other people in the way that a parent happily takes on a positive responsibility to promote good outcomes for his or her children. To paraphrase how Thomas Metzinger put things in a recent interview that he did about AI and ethics, "big tech companies typically care about promoting their own corporate interests, not about promoting the overall good" (Wendland and Metzinger 2020).

I bring up these hypothesized asymmetries in people's and organization's willingness to fill responsibility gaps by taking responsibility partly because they strike me as fascinating in themselves. But I also bring them up to illustrate that trying to fill responsibility gaps by having people volunteer themselves as responsible agents mainly seems like a realistic solution in one kind of case, namely, in the case of backward-looking positive responsibility gaps. In such cases, however, where people wish to take credit for good outcomes, it might often seem inappropriate to praise people who want to fill these responsibility gaps. The problem is that those people may not appear to truly deserve credit for the good outcomes they are willing to claim responsibility for.

10.6 Control Problems and Different Conceptions of Meaningful Human Control

In the section above, the question discussed was whether people are likely to be willing to take responsibility and thereby fill responsibility gaps. I raised the worry that in those cases that people are most likely to be willing to take responsibility – namely, cases of backward-looking positive responsibility gaps – there might be worries about whether people really deserve credit for things they might want to take responsibility for. One thing that might undermine people's claim to responsibility for some good outcome generated by an AI system might be that they did not fully understand or even predict that this was going to be the outcome. Another thing that might undermine people's claim to responsibility is that they lacked the right kind of control over the AI system. In this section, I will discuss the issue of control in particular for two reasons. Firstly, some suggestions in the literature about how responsibility gaps can be filled are about indirect forms of control, sometimes called "meaningful human control." Secondly, I brought up the issue of value alignment above when I talked about forward-looking responsibility gaps, and the issue of value alignment of AI is often discussed in relation to what is called the "control problem" regarding AI.

In my own recent work, I have discussed how to fill potential responsibility gaps associated with self-driving cars and military robots (Nyholm 2018, 2020). I have been particularly interested in responsibility gaps related to the functional autonomy of these technologies, i.e., their capacity to operate for at least certain periods of time outside of direct human control, without direct human steering. This can make some of the things these technologies do hard to predict, and together with the lack of direct control, this can seem to open up possible responsibility gaps (Hevelke and Nida-Rümelin 2015). As far as I can tell, these could be responsibility gaps of all the four kinds I have identified above. But the kind I was most interested in in my previous work was backward-looking negative responsibility gaps (e.g., ones related to harms or deaths caused by self-driving cars or military robots).

My suggestion was that we should think of these technologies as being part of human-machine teams, where the humans involved play the role of managers or supervisors, and where they can have the sort of responsibility that one can have when one is in a leading position within some form of hierarchical team (Nyholm 2018, 2020). In the military, for example, commanders can have command responsibility when a military unit is performing some operation, and they might be responsible for what others in the unit – e.g., some of the soldiers – are doing. If a military robot becomes part of the unit and can operate with some functional autonomy, the commander might also be responsible for some of what it does, just like the commander might be responsible for some of what the soldiers do.

I suggested a set of questions we can ask to home in on who should be taken to bear (the most) responsibility for what an AI-driven autonomous system is doing (Nyholm 2018, 2020). We can ask things such as:

- Under whose supervision and more or less direct control is a technology that is currently operating in "autopilot" or "autonomous" mode operating?
- Who is currently able to start, take over operation of, or, at least, stop the technology?

- Whose preferences regarding its mode of operation is the technology conforming to while in "autopilot" or "autonomous" mode?
- Who is better situated to observe and monitor the technology's actual behavior when it is in active mode?
- Who has an understanding of the functioning of the technology, at least on a "macro-level"?
- Who is able to update, or request updates of, the technology?

My idea was that the more that such questions could be answered in a way that points in the direction of a single agent or team of agents, the more it makes sense to view that agent or team of agents as being responsible for the way in which the technology system under consideration is operating. For example, a self-driving car or military robot might operate under the supervision of a certain agent, who might be able to start and stop the technology, and to whose preferences the technology's behavior might conform. That agent or team of agents might also be able to observe the operation of the technology and have an understanding, at least on a macro-level, of how the technology works. And that agent or team of agents might be able to update, or request updates to, the technology. If these conditions hold, or a significant number of them do, the agent or team of agents might justifiably be deemed responsible, I suggested, and any apparent responsibility gap might thereby be filled.

In related work, Filippo Santoni de Sio, Jeroen van den Hoven, and Giulio Mecacci have suggested a theory of what they call "meaningful human control" that has significant overlap with what I suggested (Santoni de Sio and Mecacci 2021; Santoni de Sio and van den Hoven 2018). They suggest what they call a "track and trace" theory of meaningful human control, which they think can help to fill responsibility gaps. The tracking condition states that AI-driven autonomous systems should "track" or conform to human reasons, by which I take them to mean that the system should behave in a way that fits with how the humans in question have reasons to want the systems to behave. The tracing condition states that the system should be such that there is at least some person or set of persons who have an understanding of how the system works as well as an understanding of the moral significance of having such a system in operation. The functioning of the system and its significance should "trace" back to the understanding of some such agent or group of agents. As I understand this theory, the tracking condition is similar to my question in the checklist above concerning whose preferences a technology is conforming to in its way of functioning. And the tracing condition seems to be similar to the last question on my checklist of questions that is about who has an understanding, at least on a macro-level, of how the technology works.

Notably, in their most recent paper on this, Santoni de Sio and Mecacci (2021) state their tracking condition in a language that seems to equate what they call "tracking" with what is usually called "value alignment." They write:

> Tracking requires the alignment of the system with the values, reasons, intentions of the relevant agents; tracing requires the alignment of the system with the capacities of the relevant human agents.
> (Santoni de Sio and Mecacci 2021, section 4.1)

In particular, the first part of that statement suggests that there is virtually no difference between what Santoni de Sio and Mecacci refer to as "tracking" and what is often referred to as "value alignment" of AI. Similar remarks can be made about my suggested question that asks: "whose preferences regarding its mode of operation is the technology conforming to while in 'autopilot' or 'autonomous' mode?" This also overlaps with part of at least one way of understanding the idea of value alignment, namely, that which Stuart Russell (2019) defends in his recent book *Human Compatible*. In that book, Russell understands value alignment as occurring when AI systems are conforming to human preferences.

Do the sorts of suggestions about how to resolve responsibility gaps that I myself have defended in previous work and that Santoni de Sio, van den Hoven, and Mecacci defend in their recent work help to eliminate worries about responsibility gaps? What worries remain? And what does the observation that what we have been talking about seems to be similar to what has been called "value alignment" of AI mean for how promising the above-mentioned attempted solutions are? I will now argue that we should not be too optimistic about whether the four kinds of responsibility gaps have been filled or can easily be filled.

10.7 Worries about Responsibility Gaps Revisited

I would like to suggest that the sorts of solutions to responsibility gap worries that I have sketched above – both the attempted solution I have previously defended myself and the similar one defended by Santoni de Sio et al. – are best seen as a form of idealization about how these responsibility gap worries can be dealt with in theory, rather than as practical solutions that permit smooth real-world application. Concerning the solution that I myself have suggested, for example, I already noted in my first publication on this that it could happen that when we answer the sorts of questions that I listed above, the answers might point us in different directions (Nyholm 2018, 1214). Roos de Jong (2020) develops that worry in greater detail in a response to my 2018 paper. The worry is, one might say, an updated version of the "many hands" problem.

It might be, for example, that different humans are involved, that they fulfill different elements of the criteria that I gesture toward with the questions I have sketched, and that these different humans are not all part of the "same team," so to speak.[5] For example, somebody riding in a self-driving car might be in the best position to monitor what the car is currently doing, and to some extent the behavior of the car might conform to the preferences of this person: they might want to go, say, to the grocery store, and the car might be taking them in that direction. At the same time, the person riding in the car may lack any very good understanding of how the AI in the car works – something that the engineers who designed the car have a much better understanding of. The company selling the car – or who are renting out the car to the user – might in turn be able to update the car. In other words, different people involved might all relate to the technology in question in a way that seems to make them responsible to some extent for what the technology is doing. But none of them might have enough influence and control or understanding over the technology to be viewed as the most – or most obviously – responsible for what the technology (e.g., the self-driving car) is doing (Danaher and Nyholm 2020).

The same could be the case with respect to the two conditions that are part of the "track and trace" theory that Santoni de Sio and company defend. What if the AI system (or whatever kind of technology it is) is tracking the values or reasons of one person or team of persons, but some other person or team of persons are the ones who have an understanding of the technology in question and its moral significance? And what if those two different agents or teams of agents are by no means "on the same team," so to speak? For example, perhaps they would refuse to take shared responsibility for what the technology in question is doing. Here, too, there seems to be a problem of many hands that is not easily solved.

Another worry about these views – insofar as they seem to assume that responsibility and meaningful human control require value alignment – is that achieving value alignment is widely thought to not necessarily be easy and straightforward (Gabriel 2020). Many things can go wrong. There are many potential problems. One problem is that of whose values, goals, or interests AI systems should align to. Another is that people might differ in their values, goals, or interests. This raises the question of whether the AI system should "pick sides," so to speak, or who is able to make a power move, and conform the AI system to their values, goals, and interests, which might mean that it counteracts the values, goals, and interests of others. Moreover, even if there is agreement about what values, goals, and interests ought to be tracked, problems remain. One problem is specifying these values, goals, or interests in a way that enables the AI system to reliably track them. Another is making sure that the AI system does not track these values, goals, and interests in ways that are detrimental to other important values, goals, and interests (Bostrom 2014; Russell 2019).

Furthermore, as I have argued above, there might be a forward-looking positive responsibility gap related to the goal of making sure that value alignment happens in the first place. As a reminder, I have suggested that making sure that value alignment happens is a responsibility to make sure that a good outcome is achieved. And like I noted above, both commonsense morality and many influential moral theories regard our responsibilities to positively promote good outcomes as being much weaker than our negative responsibility to prevent bad outcomes. So, there might be a responsibility gap related to the thing that is supposed to fill the other forms of responsibility gaps. In other words, if filling, for example, backward-looking negative responsibility gaps is supposed to depend on the achievement of AI value alignment, and there is a forward-looking positive responsibility gap related to who should have the responsibility of securing value alignment, then we seem to potentially have a problem.

10.8 Concluding Remarks

Let's return to AlphaGo and Lee Sedol. In that case, as was noted in the introduction, when this computer program defeated the human being in the ancient game of Go, the humans behind the computer program did not quite understand the strategies AlphaGo was using. Nor may they have fully understood, or even thought about, the moral and social significance of what they had achieved. So, what Santoni de Sio and colleagues call the "tracing" condition on meaningful human control over AI may not have been fulfilled in this case, or at least not completely fulfilled. Was the AI system "tracking" the reasons of the human beings involved? Was it aligning with human values, goals, or interests? Perhaps partly, but not necessarily wholly. Notably, Lee Sedol decided to retire from participating in Go competitions after having been defeated by AlphaGo, despite being the 18-time world champion. Competing in the game of Go lost its charm for this most accomplished of all human Go-players. In this particular case, it might be thought that by de-motivating the most accomplished human Go-player, AlphaGo did not really align with, or track, the human values or reasons related to the game of Go. So, if we assess that particular example using the "track and trace" theory from Santoni de Sio and colleagues, it might seem like there was either no meaningful human control or no meaningful value alignment, but rather a case of a lack of meaningful human control and an undermining of the values in this particular domain.

Indeed, when Carissa Veliz (2019) wrote a reflection piece on this event, she noted that a natural response to this particular case seemed to not be a sense of victory – yay, we have created an AI system that can beat the human world champion of Go! – but rather a sense of sadness. Some people may not agree with Veliz's take on this issue; they may feel that there is a very important achievement that has been made here. But nevertheless, it is not exactly clear whose achievement it was or who can take responsibility for the outcome.

Consider next the other example I started out with: the tragic death of Elaine Herzberg, who was hit and killed by a self-driving car whose AI was seemingly not doing its job, thereby failing to align with the relevant human values, goals, and interests. No human being had values, goals, and interests – it seems safe to say – that were being promoted by the hitting and killing of Elaine Herzberg by this sub-optimally performing self-driving car. In this case, it might be easier to find people who had a good understanding of how the AI system works. And it might be easier to find somebody who is able to update, or request updates to, the AI system. But those people – viz. the engineers and company behind this piece of technology – might have been different from the person best placed to monitor the performance of the technology and stop it, namely, the safety driver. The safety driver, in this case, may not have been paying sufficient attention – there is some dispute about this. So, in terms of finding who is responsible or most responsible, it might be thought that the sorts of questions I have formulated in previous work and mentioned above as well point us in different directions (Danaher and Nyholm 2020; De Jong 2020).

My conclusion, then, is that it is harder to solve real-world responsibility gap cases than it is to formulate abstract philosophical theories about the sorts of things that should be taken to bear on who is responsible for what advanced technologies like AI systems do. In some cases, especially when there is little to lose and much to gain, there might be people who willingly put themselves forward as candidates to fill responsibility gaps. But there will be many more kinds of cases where most involved parties will be unwilling to step forward and volunteer themselves as candidates to fill responsibility gaps. As noted above, filling responsibility gaps can be costly and burdensome, and so not very attractive to those we might think are good candidates.

Whether or not I have succeeded in convincing the reader that filling responsibility gaps remains a difficult problem, I hope to have convinced the reader that it makes sense to distinguish among the four general classes of responsibility gaps that I have defined in this chapter. Since it is fairly standard to distinguish between backward-looking and forward-looking responsibility, and between negative and positive responsibility, I submit that it should be undeniable that we can – and should! – distinguish among the four types of potential responsibility gaps that I have identified. That is, we should agree that there could – at least in theory – be backward-looking negative responsibility gaps, backward-looking positive responsibility gaps, forward-looking negative responsibility gaps, and forward-looking positive responsibility gaps. I hope the reader will also be inclined to agree with me that the problem of achieving value alignment of future AI systems might be a key case where there is a threat of what I have been calling a forward-looking positive responsibility gap.[6]

Notes

[1] It might be suggested here that the company Google DeepMind could take credit for having created a computer program that was able to come up with strategies that could be used to beat the world champion of Go. Does this not mean, it might be added, that they can also take credit for the victories of Lee Sedol? Not necessarily, their being able to take credit for having created such a computer program does not yet settle the question of who exactly can take credit for AlphaGo's victories over Lee Sedol in four of the five games. This can be motivated by the consideration that AlphaGo "prepared" for these games by being trained in two kinds of ways: the computer program was made to observe thousands of actual Go games and it also played millions of Go games against itself. This training led AlphaGo to generate the strategies it used in the games. The action of creating such a computer program (for which DeepMind can take credit) is different from the action of making recommendations for how to win Go games against Lee Sedol (which is something for which DeepMind might not be able to take credit, since they did not perform that action).

[2] See also p. 54 of Bonnefon et al. (2020), for another division among five different responsibility gaps, similar to those in Santoni de Sio and Mecacci's (2021) paper.

[3] According to Persson and Savulescu (2012), this is an attitude that can be explained using evolutionary psychology. I take no stance on that issue about the origins of these attitudes here.

[4] One possible suggestion here is that we reinterpret the aim of creating AI that does good by thinking of it as a way of avoiding harm created by the AI system. In other words, if we manage to create AI systems that do good, then we in effect create AI systems that avoid doing harm, and it can be seen as being clear that any developer of AI systems would have a forward-looking negative responsibility to avoid creating harmful AI systems. One striking thing about this idea, however, is that it turns the aim of creating AI that does good into a means to another end, namely, the end of avoiding harm. So, the real responsibility here, on this way of seeing things, is the responsibility to guard against risks of harm. But this seems unsatisfactory since the aim of doing good seems like it should not only be a means to other ends, but an end in itself – and it seems like there should be somebody whose positive responsibility it is to adopt that end. By translating the goal of creating AI that does good into a means for how we can avoid creating AI that does harm, we seem to lose track of the idea that it makes sense to think that there could be a positive forward-looking responsibility to create AI systems that do good. That latter responsibility is the one where common-sense morality and some important

moral theories seem to fall short when it comes to explaining who would have that responsibility. Those ways of thinking seem to make this optional – or into something that is regarded as supererogatory – rather than something that is, or should be treated as, a positive responsibility.

[5] In Danaher and Nyholm (2020), Danaher and I discuss this issue in relation to the Elaine Herzberg case mentioned in the introduction above. That can be seen as a case in point where the different people involved may have fulfilled different criteria of responsibility encapsulated in the questions I suggest that we should be asking about people interacting with AI systems.

[6] Many thanks to the editors of this volume, Adriana Placani and Stearns Broadhead, for their feedback. Thanks also to Hannah Altehenger, Susanne Burri, John Danaher, Maximilian Kiener, Sebastian Köhler, Peter Königs, Leo Menges, Daniel Tigard, Ilse Verdiesen, and an audience at the Technical University of Delft for helpful feedback. My work on this article is part of the research program Ethics of Socially Disruptive Technologies, which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

References

Bentham, Jeremy. 1789/1996. *An Introduction to the Principles of Morals and Legislation*. Oxford: Oxford University Press.

Bonnefon, Jean-Francois et al. 2020. "Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility." *European Commission*. https://op.europa.eu/en/publication-detail/-/publication/89624e2c-f98c-11ea-b44f-01aa75ed71a1/language-en

Bostrom, Nick. 2014. *Superintelligence*: *Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Danaher, John. 2016. "Robots, Law, and the Retribution Gap." *Ethics and Information Technology* 18 (4): 299–309.

Danaher, John, and Sven Nyholm. 2020. "Episode 78 – Humans and Robots: Ethics, Agency and Anthropomorphism." *Philosophical Disquisitions Podcast*: https://philosophicaldisquisitions.blogspot.com/2020/07/78-humans-and-robots-ethics-agency-and.html

Danaher, John, and Sven Nyholm. 2021. "Automation, Work, and the Achievement Gap." *AI and Ethics* 1 (2): 227–37.

Darwall, Stephen. 2004. *The Second Person Standpoint*. Cambridge: Harvard University Press.

De Jong, Roos. 2020. "The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm." *Science and Engineering Ethics* 26 (2): 727–35.

Driver, Julia. 2006. *Ethics, The Fundamentals*. Oxford: Blackwell.

Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30 (3): 411–37.

Gibbard, Allan. 1991. *Wise Choices, Apt Feelings*. New York: Oxford University Press.

Goodpaster, Kenneth. 2007. *Conscience and Corporate Culture*. Oxford: Blackwell.

Hevelke, Alexander, and Julian Nida-Rümelin. 2015. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis." *Science and Engineering Ethics* 21 (3): 619–30.

Kant, Immanuel. 1785/2012. *Immanuel Kant: Groundwork of the Metaphysics of Morals*, A German-English Edition. Cambridge: Cambridge University Press.

Keeling, Geoff. 2020. *Ethics of Automated Vehicles*. PhD thesis. University of Bristol.

Levin, Sam, and Julia C. Wong. 2018. "Self-Driving Uber Kills Ari- zona Woman in First Fatal Crash involving Pedestrian," *The Guardian*, https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe (Accessed on April 28, 2021).

Matthias, Andreas. 2004. "The Responsibility Gap." *Ethics and Information Technology* 6 (3): 175–83.

Mill, John Stuart. 1859/2012. *On Liberty*. Cambridge: Cambridge University Press.

Mill, John Stuart. 1863/2001. *Utilitarianism, Second Edition*. Indianapolis: Hackett.

Müller, Vincent. 2021. "Ethics for Machines? A Provocation." *AI4EU Beta 2:* https://www.ai4europe.eu/ethics/articles/ethics-machines-provocation

Nihlen Fahlquist, Jessica. 2017. "Responsibility Analysis." In *Ethics of Technology*, edited by S.O. Hansson, 129–42. London: Rowman & Littlefield International.

Nyholm, Sven. 2018. "Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24 (4): 1201–19.

Nyholm, Sven. 2020. *Humans and Robots: Ethics, Agency, and Anthropomorphism*. London: Rowman & Littlefield International.

Persson, Ingmar, and Julian Savulescu. 2012. *Unfit for the Future*. Oxford: Oxford University Press.

Pettit, Philip. 2007. "Responsibility Incorporated." *Ethics* 117 (2): 171–201.

Pettit, Philip. 2015. *The Robust Demands of the Good*. Oxford: Oxford University Press.

Portmore, Douglas. 2011. *Commonsense Consequentialism*. Oxford: Oxford University Press.

Robbins, Scott. 2019. "A Misdirected Principle with a Catch: Explicability for AI." *Minds and Machines* 29 (4): 495–514.

Russell, Stuart. 2019. *Human Compatible*. London: Penguin.

Russell, Stuart. 2020. "Artificial Intelligence: A Binary Approach." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 327–41. Oxford: Oxford University Press.

Santoni de Sio, Filippo, and Giulio Mecacci. 2021. "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them." *Philosophy & Technology* 34: 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Santoni de Sio, Filippo, and Jeroen van den Hoven. 2018. "Meaningful Human Control Over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI*, 5, 15. https://doi.org/10.3389/frobt.2018.00015

Searle, John. 1990. "Is the Brain's Mind a Computer Program?" *Scientific American* 262 (1): 25–31.

Sparrow, Robert. 2004. "The Turing Triage Test." *Ethics and Information Technology* 6: 203–13.

Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77.

Stilgoe, Jack. 2019. *Who's Driving Innovation?* London: Palgrave.

Tigard, Daniel. 2020. "There Is No Techno-Moral Responsibility Gap." *Philosophy & Technology* 34 (3): 589–607.

Van de Poel, Ibo, Lamber Royakkers, and Sjoerd Zwart. 2015. *Moral Responsibility and the Problem of Many Hands*. London: Routledge.

Veliz, Carissa. 2019. "A Sad Victory." *Practical Ethics: Ethics in the News*, https://blog.practicalethics.ox.ac.uk/2019/10/a-sad-victory/?fbclid=IwAR1CIVsJI3qm4fW-TmNOjN67x7UR_fNJkR1akVqS7DZC804pzqUhhBMBpgQ (Accessed on April 28, 2021).

Verdiesen, Ilse, Filippo Santoni de Sio, and Virginia Dignum. 2020. "Accountability and Control over Autonomous Weapons Systems: A Framework for Comprehensive Human Oversight." *Minds and Machines*, http://doi.org/10.1007/s11023-020-09532-9

Wendland, Karsten, and Thomas Metzinger (2020): "#05 Von Kühlschranklichtern, KI-Pubertät und Turnschuhen. Im Gespräch mit Thomas Metzinger." *Selbstbewusste KI*, http://doi.org/10.5445/IR/1000124512

Williams, Bernard, and John J. Smart. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

11 Radioactive Waste and Responsibility toward Future Generations

Céline Kermisch and Christophe Depaus

11.1 Introduction

Not only nuclear power but also medical, industrial, defense, and research activities involving ionizing radiation provide tremendous benefits for society but at the same time they cause important risks. Among these, they generate radioactive waste, some of which is potentially harmful for an extremely long period of time – for several hundred thousand years up to one million years.

The unparalleled time frame at stake raises the issue of responsibility toward future generations in a new way and requires innovative policy and technical responses to address it. The aim of this chapter is to present and analyze these strategies.

Therefore, the institutional responses from the International Atomic Energy Agency (IAEA), the Nuclear Energy Agency (NEA), the International Commission on Radiological Protection (ICRP) and the European Union are analyzed in order to clarify how these agencies frame our responsibility toward future generations. The technical responses to these new challenges are also presented. In this respect, we debate deep geological disposal – the technical solution on which there is an international consensus – and retrievability in the light of their contribution to our responsibility toward future generations, and we show that the role of retrievability is ambiguous in that respect.

It should be noted at the outset that although much philosophical analysis of future generations has been done, this chapter eschews these normative questions to focus primarily on innovative policy and technical responses. We do, however, consider certain normative dimensions associated with the specific technical issues at stake.

The chapter is structured as follows. Section 11.2 briefly reviews the specific features of radioactive waste. In Section 11.3, the institutional responses deployed to frame our responsibility toward future generations in the face of long-lived radioactive waste are introduced. In Section 11.4,

DOI: 10.4324/9781003276029-15

the technical response to these institutional requirements – the disposal of radioactive waste in a deep, stable geological formation – is presented. Section 11.5 analyzes the implementation of retrievability in the light of our responsibility toward future generations. Section 11.6 concludes this chapter.

11.2 Radioactive Waste

Radioactive waste can be defined as "material for which no further use is foreseen [...] that contains, or is contaminated with, radionuclides at concentrations or activities greater than clearance levels as established by the regulatory body" (IAEA 2007). This kind of waste mostly comes from nuclear power. However, other sources also produce nuclear waste, e.g., research, nuclear medicine, industrial activities, such as sterilization of food or medical devices by irradiation, or military and defense programs.

The picture is complex given that there are very different types of waste depending on the activity of the radionuclides contained in them and on their half-lives – the time taken for half of its atoms to decay, and thus for it to lose half of its radioactivity. The latter determines the timescale over which the waste remains radioactive. In this respect, long-lived waste is compared to short lived waste: long-lived waste refers to waste containing significant levels of radionuclides with a half-life greater than 30 years in contrast to short lived waste (IAEA 2007).

To be more precise, the IAEA distinguishes between the following waste categories: exempt waste, very short-lived waste, very low level waste, low level waste, intermediate level waste and high level waste (for a detailed presentation, see IAEA 2009a). Exempt waste, very short lived waste and very low level waste do not require exceptional handling.

Low level waste, composed of lightly contaminated items, such as tools and work clothing, includes only limited amounts of long lived radionuclides. It needs to be contained and isolated from humans and the environment for periods up to a few hundred years. Low level waste is to be disposed of in engineered surface or subsurface facilities that will be monitored and maintained as long as necessary.

Intermediate level waste contains some long lived radionuclides, and its risk ranges from several tens to several hundreds of thousands of years (ONDRAF 2011). Altogether, low and intermediate level waste represent approximately 95% of the total volume of radioactive waste, but less than 5% of the total radioactivity present in the waste (IAEA 2018). High level waste, accounting for 95% of the total activity of all waste, corresponds to waste that contains large amounts of long lived radionuclides and to waste for which the levels of activity concentration are so high that it generates significant quantities of heat (IAEA 2009a). High level waste includes mostly vitrified waste resulting from reprocessing of commercial irradiated fuel and spent fuel declared as waste. Similar to intermediate level waste, high level waste imposes a risk to humans and the environment, ranging from several tens to several hundreds of thousands of years and up to one million years. Our focus will be on these types of waste – intermediate and high level waste – given that they are the ones that raise specific challenges in terms of responsibility toward future generations.

In contrast to low level waste, the unparalleled timeframe at stake with intermediate and high level waste – far beyond any human-made project – make their management much more difficult, and they require a stronger isolation and confinement strategy. Having reviewed the types of waste and, importantly, having shown their associated timeframes, we next examine the institutional definitions of these requirements.

11.3 Responsibility toward Future Generations: Institutional Responses

Our responsibility toward future generations with respect to radioactive waste management has been framed by several institutions, most notably by the IAEA, the NEA, the International Commission on Radiological Protection and the European Union. These institutional responses are presented below.

11.3.1 The Radioactive Waste Management Principles of the IAEA

In 1995, the IAEA established "radioactive waste management principles" conceived to guide choices to be made in radioactive waste management (IAEA 1995). Among these, principles 4 and 5 specifically address the issue of our responsibility toward future generations.

> Principle 4: Protection of future generations
>
> Radioactive waste shall be managed in such a way that predicted impacts on the health of future generations will not be greater than relevant levels of impact that are acceptable today [...]
>
> This principle is derived from an ethical concern for the health of future generations [...] the intent is to achieve reasonable assurance that there will be no unacceptable impacts on human health [...].
>
> Principle 5: Burdens on future generations
>
> Radioactive waste shall be managed in such a way that will not impose undue burdens on future generations [...].

> This principle is based on the ethical consideration that the generations that receive the benefits of a practice should bear the responsibility to manage the resulting waste [...]. The responsibility of the present generation includes developing the technology, constructing and operating facilities, and providing a funding system, sufficient controls and plans for the management of radioactive waste [...]. The identity, location and inventory of a radioactive waste disposal facility should be appropriately recorded and the records maintained".
>
> (IAEA 1995)

These two principles correspond to future-oriented constraints for radioactive waste management (Kermisch 2021). They refer to intergenerational equity – which is justice between generations. Indeed, on the one hand, principle 4 requires future generations to be as well protected as current generations. On the other hand, principle 5 requires generations who have benefited from nuclear energy to manage the resulting waste so to avoid undue burdens on future generations who would not have benefited from nuclear activities, and to transfer the knowledge of the waste to these future generations.

Let us mention furthermore that principle 6, which refers to the national legal framework, also considers responsibilities toward future generations:

> Principle 6: National legal framework [...]
>
> Since radioactive waste management can span timescales involving a number of human generations, appropriate consideration of present and likely future operations should be taken into account. Provisions for sufficiently long lasting continuity of responsibilities and funding requirements should be made.
>
> (IAEA 1995)

11.3.2 The Ethical Basis of Geological Disposal of Radioactive Waste of the NEA

In the same vein, the NEA of the Organization for economic co-operation and development (OECD) has established *the environmental and ethical basis of geological disposal of long lived radioactive wastes*, to be used in making choices about the waste management strategy to be adopted (NEA 1995). It also tackles the issue of our responsibility toward future generations:

- the liabilities of waste management should be considered when undertaking new projects;

- those who generate the waste should take responsibility and provide the resources, for the management of these materials in a way which will not impose undue burdens on future generations;
- waste should be managed in a way that secures an acceptable level of protection for human health and the environment and affords to future generations at least the level of safety which is acceptable today; there seems to be no ethical basis for discounting future health and environmental damage risks;
- a waste management strategy should not be based on a presumption of a stable societal structure for the indefinite future, nor of technological advance; rather it should aim at bequeathing a passively safe situation which places no reliance on active institutional controls (NEA 1995).

The ethical basis of radioactive waste management advocated by the NEA coincides with the IAEA principles with respect to the requirement of equal protection for future generations – given that discounting the value of future generations is not justified – and in avoiding undue burdens on them.

11.3.3 The Ethical Foundations of the System of Radiological Protection of the ICRP

ICRP Publication 138 is dedicated to the ethical foundations of the system of radiological protection elaborated within the previous ICRP publications (ICRP 2018). It unveils the core ethical values underpinning the system of radiological protection, namely beneficence and non-maleficence, prudence, justice, dignity but also the procedural values of accountability, transparency, and inclusiveness (stakeholder participation).

Two paragraphs of ICRP Publication 138 are dedicated to our responsibilities toward future generations.

Firstly, paragraph 58 addresses the issue of intergenerational distributive justice.

> (58) Intergenerational distributive justice has been addressed by the Commission for the management of radioactive waste with reference to 'precautionary principle and sustainable development in order to preserve the health and environment of future generations' (ICRP, 2013, Para. 15). In Publication 81, the Commission recommends that 'individuals and populations in the future should be afforded at least the same level of protection as the current generation' (ICRP, 1998, Para. 40). In Publication 122, the Commission introduces responsibilities towards future generations in terms of providing the means to deal with their protection: '... the obligations of the present

> generation towards the future generation are complex, involving, for instance, not only issues of safety and protection but also transfer of knowledge and resources' (ICRP, 2013, Para. 17).
>
> (ICRP 2018)

Secondly, paragraph 68 addresses accountability of the present generation to future generations.

> (68) The Commission also considered the accountability of the present generation to future generations, which is mentioned explicitly in Publications 77 (ICRP, 1997), 81 (ICRP, 1998), 91 (ICRP, 2003), and 122 (ICRP, 2013) related to waste management and the protection of the environment. As an example, Publication 122 (Para. 17) states '[...] Due to the technical and scientific uncertainties, and the evolution of society in the long term, it is generally acknowledged that the present generation is not able to ensure that societal action will be taken in the future, but needs to provide the means for future generations to cope with these issues' (ICRP, 2013). Accountability in this context is part of implementing the value of intergenerational distributive justice.
>
> (ICRP 2018, 35)

As with the IAEA and the NEA, the ICRP expresses a requirement of equal protection for future generations as part of our obligations in terms of intergenerational justice. Besides, when considering the protection of future generations, the commission highlights the fact that not only safety is at stake, but also transfer of knowledge and resources – as also mentioned by the IAEA (IAEA 1995).

11.3.4 The Joint Convention

The *Joint Convention on the Safety of Spent Fuel Management and on the Safety of Radioactive Waste Management* was adopted in December 1997 (IAEA 1997). It is the only legally binding instrument that addresses the issue of spent fuel and radioactive waste management safety on a global scale.

In Article 1, the *Joint Convention* emphasizes a duty of sustainability for future generations:

> The objectives of this Convention are: [...] (ii) to ensure that during all stages of spent fuel and radioactive waste management there are effective defences against potential hazards so that individuals, society and the environment are protected from harmful effects of

> ionising radiation, now and in the future, in such a way that the needs and aspirations of the present generation are met without compromising the ability of future generations to meet their needs and aspirations; [...].
>
> (IAEA 1997)

Article 11 defines further requirements for future generations:

> [...] each Contracting Party shall take the appropriate steps to: [...] (vi) strive to avoid actions that impose reasonably predictable impacts on future generations greater than those permitted for the current generation; (vii) aim to avoid imposing undue burdens on future generations.
>
> (IAEA 1997)

The focus of the *Joint Convention* is in line with the requirements identified above: safety should be ensured through equality of protection and avoidance of undue burdens on future generations.

11.3.5 The European Union Waste Directive

At the European level, the Council of the European Union has adopted Directive 2011/70/Euratom in 2011 (EU 2011).

In the whereas clauses, the same commitment to avoid undue burdens on future generations is reiterated:

> (24) It should be an ethical obligation of each Member State to avoid any undue burden on future generations in respect of spent fuel and radioactive waste including any radioactive waste expected from decommissioning of existing nuclear installations. Through the implementation of this Directive Member States will have demonstrated that they have taken reasonable steps to ensure that that objective is met.
>
> (EU 2011)

11.3.6 An Institutional Framework Oriented toward Passive Safety

As we can see from the IAEA and the NEA principles, the ICRP and the European Union, two recurrent challenges in terms of responsibilities toward future generations emerge when dealing with intermediate and high-level wastes.

On the one hand, we must keep the same level of protection for future generations as for current generations. Let us note that this requirement itself has been widely debated not only on the basis of its feasibility given the considerable uncertainties at stake, but also on the basis of its ethical desirability (Okrent 1999; Shrader-Frechette 1993; Taebi 2012). It is indeed uncertain whether we are allowed to impose similar risks on future generations as on ours, since we are the ones who have benefitted from nuclear power.

On the other hand, the generations who have benefitted from nuclear energy must take care of the waste that they have generated in order to avoid undue burdens on future generations who might not benefit from nuclear energy anymore.

In the latter respect, the core ethical concept that is advocated for ensuring our responsibility toward future generations for long-lived radioactive waste management is the concept of *passive safety*. The management option to be chosen must be "passively safe" in the long term, meaning that the system should not require any human intervention to guarantee its safety. Indeed, at these time scales, it is impossible to guarantee societal stability and thus to ensure that future people will be able to take care of the waste – they could lose the necessary knowledge or lack adequate means to do so, for instance. Hence, the need to find a solution that does not necessitate any human intervention in the long run. In other words, the chosen option must not require the involvement of future generations to maintain the system safely and thereby not impose any burden on them. In this respect, deep geological disposal is the reference option at the international level for the management of intermediate and high-level wastes.

11.4 Responsibility toward Future Generations: Deep Geological Disposal as the Technical Solution

The principle of deep geological disposal is based on the containment of radioactive wastes and on their isolation from the biosphere in a deep, stable geological host formation until their level of radioactivity becomes comparable to that of natural radioactivity – which takes several hundred thousand years for long-lived waste.

As of 2022, several countries are in the process of implementing this kind of facility for the handling of civil radioactive waste (Finland, Sweden, France, Germany, Russia, etc.). Regarding military radioactive waste, the Waste Isolation Pilot Plant (WIPP) in New Mexico has hosted waste resulting from the research and production of the United States since 1999.

Practically, the disposal of radioactive waste in a deep geological disposal relies on the fact not only that the waste is contained in human-made packages, but also – and most importantly – that these packages are disposed of deep underground, at a depth of several hundred meters, in a stable geological host formation. It is thus the combination of engineered and natural barriers that ensure the safety of the disposal. Concretely, engineered barriers designate the various layers of packaging around the waste, i.e. the solid form of the radioactive waste itself, the long-lasting container, the material placed immediately around the waste containers to add further protection as well as any other engineered features of the disposal facility such as the backfill, plugs, seals in tunnels or vaults. The natural barrier refers to the geological formation in which the waste packages are emplaced. As mentioned previously, the general safety strategy of a geological disposal is based on containment and isolation of the waste (IAEA 2011). More precisely, containment refers to the properties of the system to retain radionuclides. It is ensured both by the engineered and the natural barriers. Isolation designates the physical distancing from human beings. It is mainly ensured by the host rock and its environment.

The implementation of a geological disposal is expected to last approximately 100 years, including the licensing phase, the construction phase, and the operational phase. Even though not required for safety reasons, it is generally expected to be followed by an institutional control phase to accommodate potential societal requests. After that, no human intervention is expected to be needed. The system will thus be passively safe, with long-term safety ensured both by engineered and natural barriers. Hence, in principle, no undue burdens will be imposed on future generations.

Furthermore, deep geological disposal also theoretically addresses the second requirement in terms of responsibilities toward future generations expressed at the institutional level, i.e. the fact that we must keep the same level of protection for future generations as for current generations. Indeed, the predicted impacts on the health of future generations of deep geological disposal are not expected to be greater than the ones that are acceptable today: the dose constraints for this type of facility are 0.3 mSv/year according to the ICRP (ICRP 2007, 2013) – and usually lower according to national legislations, i.e. 0.1mSv in Belgium, for instance. Such a dose corresponds only to a minor fraction of the annual dose received by an individual – where the annual doses due to natural sources of radiation in millisieverts are estimated to be 2.4 mSv/year according to the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR).[1]

Besides avoiding undue burdens on future generations and ensuring equal protection for them, it has been argued that the transfer of the knowledge of the waste to future generations must be ensured. This requirement can not only be seen as associated with safety, but it can also be considered in relation to the question of freedom of choice for future generations. The latter is one of the reasons why deep geological disposal is sometimes conceived as being reversible and/or retrievable.

11.5 The Impact of Retrievability on Deep Geological Disposal

The NEA defines reversibility as "the ability in principle to reverse decisions taken during the progressive implementation of a disposal system; reversal is the actual action of going back on (changing) a previous decision, either by changing direction, or perhaps even by restoring the situation that existed prior to that decision. Reversibility implies making provisions in order to allow reversal should it be required" (NEA 2012a). On the other hand, retrievability is defined as "the ability in principle to recover waste or entire waste packages once they have been emplaced in a repository; retrieval is the concrete action of removing the waste. Retrievability implies making provisions in order to allow retrieval should it be required" (NEA 2012a). Reversibility thus relates to the decision-making process, whereas retrievability refers to physical interventions on the waste. Our focus is on the latter. Indeed, whereas reversibility is conceived as an industry good practice that is endorsed by most countries that are in the process of implementing a geological disposal, there is no consensus about retrievability. In some countries, the possibility to retrieve the waste for a variety of reasons – safety-related, societal, economical etc. – is foreseen. This is the case in France, for instance, where the waste must be retrievable for more than 100 years. The situation is similar in Belgium and Switzerland. In the United States, the United States Department of Energy requires, in the demonstration of compliance, a report explaining how the removal of waste from the disposal system could be done for a reasonable period of time after disposal (NEA 2012b). On the other hand, some countries do not consider retrievability of the waste, except in the case of safety-related motives and for a limited duration only. This is the case in Sweden and Finland, for instance.

Practically, retrievability is not unequivocal. There are indeed various types of provisions that can be implemented in order to increase the retrievability of a geological disposal. As general guidelines, the NEA states that "at the technical level, the application of retrievability provisions will depend on such factors as the host geology, the engineered barrier concepts and the life cycle phase(s) of the repository during which retrievability is desired. [...] Examples of provisions increasing retrievability include: more durable waste forms and waste containers, longer periods granted before closing galleries and the final repository, and buffer and backfill materials that are easier to remove" (NEA 2012a). Let us note at this stage that considering longer periods before sealing the disposal delays the implementation of long-term passive safety.

From an ethical standpoint, one of the arguments invoked by philosophers and by the public in favor of retrievability is that it can be seen as a way to preserve the autonomy of future generations insofar as it leaves the possibility for them to opt for another solution to radioactive waste management if they want to do so (Andren 2012; Fondation Roi Baudouin 2010; Shrader-Frechette 1993). They could wish to do so for a variety of reasons. Indeed, future generations could want to retrieve the waste for safety reasons – in case of incident or leakage or because new, safer, technologies become available. They could also be willing to do so for reasons that are more economic in nature – for instance, they could decide to recover the waste in order to extract and recycle the fissile materials contained in the waste, which would of course also make sense from a sustainability perspective.

However, retrievability is far from being the ethical panacea. Indeed, while it contributes to keeping options open for future generations, it introduces new ethical problems and conflicts that require imposing severe restrictions on retrievability.

First, it is important to be aware of the fact that it is wrong to state that retrievability contributes to the freedom of choice of future generations in general, as if they were one monolithic group. Indeed, the idea of freedom of choice only makes sense as long as future generations have the memory of the waste and its location (Kermisch 2016). These could be defined as close future generations, as opposed to remote future generations who would have lost this memory (Kermisch and Depaus 2021). The memory issue is also strongly related to our responsibility toward future generations in terms of the transfer of knowledge, advocated by the IAEA and the ICRP (IAEA 1995; ICRP 2018). Yet, it seems unreasonable to assume that memory of the waste will be kept long enough. Indeed, intermediate and high-level wastes are harmful for several hundred thousand years and humility, at least, supports the assumption that we will not be able to transfer knowledge over such a long period. In this respect, we might note for instance that, according to the Autorité de Sûreté Nucléaire (ASN) – the French nuclear safety authority – this knowledge can be expected to be kept for no less than 500 years (ASN 2008), which is of course incomparable to the timescale involved. Hence, we can say that retrievability indeed contributes to the freedom of choice of future generations, but only of close future generations (Kermisch 2016). This observation also justifies a geological disposal being retrievable for only a limited timeframe, during which it is essential to transfer the relevant knowledge.

Second, retrievability potentially conflicts with safety. Indeed, safety can be assessed in many different ways. In the context of a qualitative assessment of retrievability, the focus on health impact appears to be relevant (Kermisch 2016; Kermisch and Depaus 2021). More specifically, safety is related to the potential exposure of individuals, defined as the "exposure that is not expected to be delivered with certainty but that may result from an accident at a source or an event or sequence of events of a probabilistic nature, including equipment failures and operating errors. Due to the large uncertainties surrounding exposures that may occur in the future, they are considered as potential exposures" (ICRP 2013). Potential exposure is determined by four dimensions (ONDRAF 2011): (1) the distance between the radiation source and the receiver, (2) the potential for harm of the source, which is a function of the radiological characteristics—the decay mode, the decay half-life, and the volumic activity, (3) the presence of protection barriers and their characterization, and (4) the likelihood of contact of receivers, associated with "planned exposure situations", corresponding to "exposure situations resulting from the operation of deliberately introduced sources" (ICRP 2013). A fifth component can be added in order to give a more exhaustive picture of the stakes at the safety level: (5) the possibility to monitor the facility and the waste and to proceed with maintenance if necessary (Kermisch 2016; Kermisch and Depaus 2021). With this conception of safety in mind, we can see that, for close future generations, retrievability positively affects the possibility to monitor and maintain the facility. On the other hand, if the galleries and shafts are kept open longer that technically necessary, retrievability negatively impacts the protection barriers for all future generations, as well as the likelihood of contact for close future generations. Furthermore, conceptually, the long-term safety of a geological disposal is ensured through passive safety, which implies that no human intervention is foreseen. The latter is intrinsically contradictory to the idea of keeping the repository retrievable (NEA 2021). For safety reasons, retrievability should thus be limited in time.

Third, retrievability also conflicts with nuclear security. By definition, nuclear security refers to "the prevention and detection of, and response to, theft, sabotage, unauthorized access, illegal transfer or other malicious acts involving nuclear material, other radioactive substances or their associated facilities" (IAEA 2007). Whereas security refers to intentional harm, safety refers to unintended harm. It appears obvious that the implementation of any provision designed to facilitate the retrievability of the waste is in principle opposed to the aim of avoiding the access to these materials – the IAEA even assumes that the waste could be diverted from a retrievable geological disposal in a few days, whereas it would take several years with a sealed geological disposal (IAEA 2009b).

Fourth, retrievability could also conflict with operational safety. Indeed, one possibility for increasing the retrievability of the waste consists in designing a facility with shorter disposal galleries. Shorter galleries facilitate the retrieval of the waste packages put in place at the back of the gallery. With the same inventory, such a choice multiplies the number of galleries and implies a more complicated facility layout. More crossings are thus needed, which could affect ultimately the safety of workers during operations.

Fifth, postponing the closure of a geological disposal after the operational phase also imposes undue burdens on future generations, which would be responsible for the funding of the monitoring and maintenance, as well as for the long-term transfer of knowledge.

In the end, the role of retrievability is ambiguous with respect to our responsibility toward future generations. On the one hand, it allows us to account for the freedom of choice of close future generations but, on the other hand, it could lead to severe safety and security deficiencies. These challenges in terms of safety and security could be alleviated by considering retrievability for a limited duration only, so to avoid postponing passive safety and prolonging the conflict about the access to radioactive materials. The temporal framing of retrievability is thus essential.

11.6 Conclusion

This chapter has shown that the management of intermediate and high-level waste raises new challenges in terms of responsibilities toward future generations. This is clearly highlighted by the specific institutional framing of radioactive waste management, which emphasizes two main duties toward future generations: ensuring an equal level of protection for them and not imposing undue burdens on them.

Deep geological disposal of radioactive waste, the technical solution on which there is international consensus, addresses these two requirements. Indeed, on the one hand, its predicted impacts on the health of future generations are not expected to be greater than the ones that are acceptable today and, on the other hand, the passive safety of this option does not require any intervention of future people in order to ensure the safety and security of the facility.

The geological disposal facility is conceived by some countries as allowing the retrievability of the waste in order to preserve the freedom of choice of future generations. However, we have highlighted that retrievability could be problematic in terms of safety and security if not framed temporally. This is specifically the case in the long run, where the advantages of retrievability are lost in any case. Hence, from the standpoint of our responsibility toward future generations, it appears that retrievability should not be implemented without restrictions.

Note

[1] https://www.unscear.org/unscear/en/areas-of-work/radiation-faq.html

References

Andren, Mats. 2012. “An Uncomfortable Responsibility: Ethics and Nuclear Waste.” *The European Legacy* 17 (1): 71–82.

ASN (Agence de sûreté nucléaire). 2008. *Guide de sûreté relatif au stockage définitif des déchets radioactifs en formation géologique profonde*. ASN. https://www.asn.fr/content/download/50883/352509?version=2

EU (European Union). 2011. Directive 2011/70/Euratom of 19 July 2011 establishing a Community framework for the responsible and safe management of spent fuel and radioactive waste, JO L 199.

Fondation Roi Baudouin. 2010. *Conférence citoyenne ‘Comment décider de la gestion à long terme des déchets radioactifs de haute activité et de longue durée de vie?’*.

IAEA (International Atomic Energy Agency). 1995. *The Principles of Radioactive Waste Management*. Safety series No 111-F. Vienna: IAEA.

IAEA (International Atomic Energy Agency). 1997. *Joint Convention on the Safety of Spent Fuel Management and on the Safety of Radioactive Waste Management*. Vienna: IAEA.

IAEA (International Atomic Energy Agency). 2007. *IAEA Safety Glossary*. 2007 ed. Vienna: IAEA.

IAEA (International Atomic Energy Agency). 2009a. *Classification of Radioactive Waste*. General safety guide No GSG-1. Vienna: IAEA.

IAEA (International Atomic Energy Agency). 2009b. *Geological Disposal of Radioactive Waste: Technological Implications for Retrievability*. Report NW-T-1.19. Vienna: IAEA.

IAEA (International Atomic Energy Agency). 2011. *Disposal of Radioactive Waste*. Specific safety requirements No SSR-5. Vienna: IAEA, 2011.

IAEA (International Atomic Energy Agency). 2018. *Status and Trends in Spent Fuel and Radioactive Waste Management*, IAEA Nuclear Energy Series No. NW-T-1.14. Vienna: IAEA.

ICRP (International Commission on Radiological Protection). 1997. “Radiological Protection Policy for the Disposal of Radioactive Waste”. ICRP Publication 77, *Annals of the ICRP* 27 (suppl.): 1–21.

ICRP (International Commission on Radiological Protection). 1998. “Radiation Protection Recommendations as Applied to the Disposal of Long-Lived Solid Radioactive Waste”. ICRP Publication 81, *Annals of the ICRP* 28(4): 1–25.

ICRP (International Commission on Radiological Protection). 2003. “A Framework for Assessing the Impact of Ionising Radiation on non-Human Species”. ICRP Publication 91, *Annals of the ICRP* 33(3): 207–63.

ICRP (International Commission on Radiological Protection). 2007. “The 2007 Recommendations of the International Commission of Radiological Protection”. ICRP Publication 103, *Annals of the ICRP* 37(2–4): 1–32.

ICRP (International Commission on Radiological Protection). 2013. “Radiological Protection in Geological Disposal of Long-Lived Solid Radioactive Waste”. ICRP Publication 122, *Annals of the ICRP* 42(3): 7–57.

ICRP (International Commission on Radiological Protection). 2018. “Ethical Foundations of the System of Radiological Protection”. ICRP Publication 138, *Annals of the ICRP* 47(1): 7–61.

Kermisch, Céline. 2016. "Specifying the Concept of Future Generations for Addressing Issues Related to High-Level Radioactive Waste." *Science and Engineering Ethics* 22 (6): 1797–811.

Kermisch, Céline. 2021. Radioactive Waste. In *The Palgrave Global Handbook of Sustainability*, edited by Robert Brinkmann. New York: Palgrave Macmillan.

Kermisch, Céline, and Christophe Depaus. 2021. "The Strength of Ethical Matrixes as a Tool for Conceptual Normative Analysis Related to Technological Choices: the Case of Geological Disposal for Radioactive Waste." *Science and Engineering Ethics* 24: 29–48.

NEA (Nuclear Energy Agency). 1995. *The Environmental and Ethical Basis of Geological Disposal of Long-Lived Radioactive Waste*. Paris: OECD.

NEA (Nuclear Energy Agency). 2012a. *Reversibility of Decisions and Retrievability of Radioactive Waste*, Report NEA 7085. Paris: OECD.

NEA (Nuclear Energy Agency). 2012b. *Reversibility and Retrievability in Planning for Geological Disposal of Radioactive Waste*, Proceedings of the "R&R" International Conference and Dialogue 14-17 December 2010, Reims, France. Report NEA 6993. Paris: OECD.

Okrent, David. 1999. "On Intergenerational Equity and Its Clash with Intragenerational Equity and on the Need for Policies to Guide the Regulation of Disposal of Wastes and Other Activities Posing Very Long-Term Risks." *Risk Analysis* 19 (5): 877–901.

ONDRAF. 2011. *Waste plan for the long-term management of conditioned high-level and/or long lived radioactive waste and overview of related issues*. Report NIROND 2011-02E.

Shrader-Frechette, K.S. 1993. *Burying Uncertainty: Risk and the Case Against Geological Disposal of Nuclear Waste*. Berkeley: University of California Press.

Taebi, Behnam. 2012. Intergenerational Risks of Nuclear Energy. In *Handbook of Risk Theory. Epistemology, Decision Theory, Ethics and Social Implications of Risk*, edited by Sabine Roeser, Rafaela Hillerbrand, Martin Peterson and Per Sandin. Dordrecht: Springer.

Part V

Environmental Context

12 Resilience and Responsibilities

Normative Resilience for Responsibility Arrangements

Neelke Doorn and Samantha Copeland

12.1 Introduction

Climate change is one of today's great societal challenges, and it is associated with a multitude of risks and high levels of uncertainty. It is therefore not surprising to see that climate change often serves as a context to illustrate different concepts related to risk. The focus of the current chapter is on resilience.

It is now widely accepted that climate change requires both mitigation actions to reduce climate change and adaptation measures to cope with the effects and increased risks brought about by climate change, such as droughts, heat waves, heavy rainfall, and flooding, among others (IPCC 2022). In recent years, resilience has emerged as one of the leading paradigms for climate adaptation policy (Doorn 2017; Fünfgeld and McEvoy 2012; Twigger-Ross et al. 2011).

After a first wave of enthusiasm in the literature, resilience is increasingly becoming a contested concept. Not only does the concept lack clarity due to theoretical inconsistencies and ambiguity in its use (Deppisch 2017), but definitions of resilience also uniformly portray resilience as a desirable goal, which is problematized by research that questions the distribution of benefits and burdens under different resilience regimes (Meerow, Newell, and Stults 2016). A growing number of scholars now recognize that, for climate adaptation to draw on and benefit in practical ways from a resilience approach, the appropriation and use of resilience to justify policy measures should be critically scrutinized, as it contains particular normative choices that are often not made explicit (Copeland et al. 2020; Cote and Nightingale 2011; McEvoy, Fünfgeld, and Bosomworth 2013; Porter and Davoudi 2012).

Although it is often said that resilience involves new responsibility arrangements between state and local actors (Butler and Pidgeon 2011), with an increasing emphasis on the responsibilities of citizens (Doorn, Brackel, and Vermeulen 2021), the literature has hitherto devoted limited

DOI: 10.4324/9781003276029-17

attention to the responsibilities that citizens are expected to assume under different resilience regimes (Hegger et al. 2017). A more prominent role for citizens cannot be a simple substitute for responsive and accountable governance (Cretney 2014; Davoudi 2012). In this chapter, we will look at resilience in the context of climate adaptation. In addition to the more general literature on resilience, we will also look specifically at the literature on urban resilience, which highlights two different, but for climate adaptation equally relevant, framings of resilience in relation to either a system or a community. We thereby draw out the implications of implicit normativity in various conceptions and critiques of resilience as a framing concept. The aim of this chapter is to develop an explicitly normative notion of resilience that can account for the responsibilities of different actors in realizing resilience. Our account provides a conceptualization that links the normative aspects of resilience and its application in the context of climate adaptation in a way that makes accompanying responsibility arrangements explicit components of resilient systems.

12.2 Origins of Resilience

Although the term resilience is in itself not new – its early use dates back to the 18th century when it was used to denote the strength of materials (McAslan 2010) – the contemporary use of the term resilience, as a concept that typically applies to *systems* rather than *isolated components*, originates from discussions in system dynamics and ecology in the 1960s (Holling 1973). Holling's paper should be seen in light of the rise of systems thinking (Davoudi 2012). Crucial for systems thinking is that the performance of a complex system comprises more than the performance of the parts or components that make up the system. Analysis of these systems should therefore not focus on the stable performance of the different components, but rather on the relations between these components and how performance emerges from interactions between these components. In ecology, resilience was used as an explanatory concept, highlighting the various processes in dynamic complex systems that produce the high degree of stability and adaptability that we observe in natural ecosystems despite facing a wide range of external perturbations and conditions (e.g., Folke 2006; Walker et al. 2006). After its introduction into ecology in the 1960s and 1970s, the term resilience became popular in other domains as well, entering the domain of safety engineering around 2000 (Woods and Wreathall 2003), psychology (Connor and Davidson 2003; Southwick, Vythilingam, and Charney 2005), disaster management (Adger 2000; Paton and Fohnston 2001), and business (Hamel and Valikangas 2003). With this, a *social* dimensions began to be integrated into resilience thinking as well (Adger et al. 2009;

Davoudi 2012), prompting contemporary approaches to take complex social-technical-environmental systems as their field of application (4TU. Resilience Engineering 2021).

In addition to the question of scope, another ongoing question in the literature is whether resilience should be seen as an outcome or as a process or ability (Cañizares, Copeland, and Doorn 2021). When seen as an outcome, the concept allows us to specify resilient things, but only after disturbance (*ex-post*). However, resilience can sometimes also be seen as an ability that enables things to display desirable response(s) to some disturbance(s). This ability is usually expressed in terms of resilience determinants that characterize resilient things before disturbance (*ex-ante*) and are key for designing or managing resilience.

Additionally, resilience concepts also differ in their normativity. Until the end of the 20th century, resilience was considered a neutral, technical characteristic of a system, which primarily referred to the ability of a system to return to an equilibrium state. However, with the adoption of the resilience paradigm in policy circles, resilience adopted an explicit normative dimension, functioning as an ideal that communities should strive for (Béné et al. 2017). This ideal is generally understood in terms of the capacity of a community to absorb disturbance and re-organize while undergoing change so as to still retain essentially the same functions (Folke et al. 2011), for example, when people support their elderly neighbors to evacuate their house in case of flooding.

12.2.1 Different Frames of Urban Resilience

Unsurprisingly, with the use of the same term in different domains, different definitions and interpretations of resilience emerged emphasizing different aspects of Holling's general idea. This is not problematic per se, as long as this polysemy is recognized, and it is clear how resilience is interpreted. Recent attempts to better articulate the resilience concept in the context of the social processes of climate adaptation have identified three discrete characteristics of resilience that could be usefully applied to adaptation (cf. Adger et al. 2011; Berkes 2017; Folke et al. 2010; Turner 2010; Walker et al. 2006):

- Resilience understood as the ability to maintain functions after disturbance (Walker et al. 2004)
- Resilience understood as a system's capacity to self-organize (Folke et al. 2004)
- Resilience as the capacity to learn and adapt (Folke 2006)

Although these three characteristics provide a distinctly positive way to discuss urban climate adaptation and other urban policy agendas

(McEvoy, Fünfgeld, and Bosomworth 2013), they do not prescribe a particular set of actions or any specific way to measure or evaluate resilience (Meerow, Newell, and Stults 2016).

In the literature on urban resilience, two dominant frames of resilience have emerged, emphasizing different characteristics of resilience (Wardekker 2022). The first is a "system framing" of resilience, which emphasizes its roots in system dynamics and which is also the most common in policy discourse (e.g., Biggs, Schlüter, and Schoon 2015; Eraydin and Taşan-Kok 2013; Martin and Sunley 2015; Sharifi and Yamagata 2016; Shutters, Muneepeerakul, and Lobo 2015). Urban resilience is, for example, "the ability of the city to maintain the functions that support the well-being of its citizens" (Da Silva, Kernaghan, and Luque 2012), conceptualizing cities as systems with components, functions, and flows of, among other things, resources, materials, and people (e.g., Meerow, Newell, and Stults 2016; Wardekker et al. 2010). This framing of resilience is outcome-oriented, with larger stakeholders and authorities often considered the natural key players (Wardekker 2022). A potential blind spot of this "system framing" is that it can fail to take note of disproportionate impacts on specific subsystems or vulnerable subpopulations and of the fact that the role of local actors (most notably, individual citizens) can be hardly accounted for in system-level descriptions (ibid).

The second is a "community framing" of urban resilience, which has its roots in disaster preparedness and psychology, and which focuses on the impact of disturbances on communities (Gunderson 2010; Norris et al. 2008). Local citizens and other small stakeholders are the key players in this framing of urban resilience, emphasizing urban life, community bonds, and self-sufficiency. Typical resilience principles are derived from social science literature, such as social networks, leadership, engagement, information flow, learning, societal partnerships, and societal equity (e.g., Berkes and Ross 2013; Brown and Westaway 2011; Chandra et al. 2010; Leichenko 2011; Zurlini et al. 2013). A potential blind spot of this framing is that it particularly focuses on shocks that directly impact communities, potentially neglecting slower, creeping stresses and interactions with other levels and scales (Wardekker 2022).

Both sub-literatures on urban resilience have thus hitherto devoted limited attention to the responsibilities that private actors are expected to assume under different resilience regimes and how these should be complemented with public actors' responsibilities (Hegger et al. 2017). The system frame focuses primarily on the role of infrastructures and not individual citizens. The community frame pays little attention to the interaction between citizens and actors, except in terms of the community itself. The approach we take in this chapter offers guiding steps (and language) that will enable fruitful deliberation in practice about how to link the actions and responsibilities of individuals who make up communities to the overall resilience of a system.

12.3 Criticism of Resilience

Before looking at how actors' responsibilities could be incorporated into resilience, let us take a closer look at some of the criticism voiced against the ideal of resilience. We will provide a brief summary of three main points of criticism found in the resilience literature. By elucidating the aspects of resilience concepts that we should avoid in the context of climate adaptation, we seek to avoid the pits into which a naïve conceptualization of resilience might fall.

Bounce back to a state where one doesn't want to be in the first place: Several authors warn against the interpretation of resilience as bounce-back (Jordan and Javernick-Will 2013; Twigger-Ross et al. 2014). Not only can bounce-back be unrealistic, but it may also lead to the reproduction of vulnerabilities and other undesirable situations. If the aim is merely to return to the previous state (what was considered to be "normal") without questioning what such normality entails (Pendall, Foster, and Cowell 2010) or whether that state is desirable, then resilience may run the risk of reproducing undesirable situations (Cannon and Müller-Mahn 2010). Hurricane Katrina is often mentioned as an example of a disaster that revealed social processes that many people did not consider an acceptable, pre-disaster situation to which they wanted to return (Davoudi 2012). Similarly, Barnett argues that recovery is insufficient in the longer term; in a context of uncertainty, a resilient system should not just bounce back but "bounce back in better shape" (Barnett 2001, 984) or "bounce forward" (Shaw and Maythorne 2013), because that will enable the system to better cope with uncertainty and deal with surprises. Ideally, a resilient system should be able to adapt and transform so that it can deal with new situations. Over longer timescales, a resilient system should "encompass the dynamics to accommodate trends and co-evolve" (Wardekker et al. 2010, 988). However, despite an increasing recognition in the academic literature that recovery is not enough, the "engineering" view of resilience as bouncing back dominates policy and resilience practice (Meerow and Stults 2016).

Relation to vulnerability research: The concept of resilience is increasingly replacing vulnerability as the focus of the literature on disaster risk reduction, which prompts questioning the relation between the two notions, as well as what is lost if resilience replaces vulnerability as the dominant paradigm. Although some authors see vulnerability and resilience as flipsides or opposites – which would render the shift from vulnerability to resilience a matter of mere rhetoric – most authors recognize that the relationship between vulnerability and resilience is more complicated than this and that the terms are used in different ways. Part of this can be traced back to the disciplinary origins of the two concepts with different associated epistemological traditions (Gallopín 2006; Janssen et al. 2006). Originating from the natural (ecological, biophysical) sciences, resilience suggests a strong positivist epistemology with a focus on objective definitions and measurements of relevant phenomena (Miller et al. 2010). Vulnerability research, by contrast, has its origins in the social sciences and has been influenced by a stronger constructivist epistemology in which the very notion of vulnerability is the product of diverse human cultures and agency, where differential vulnerability among individuals and groups may be produced even when confronting seemingly identical risks (McLaughlin and Dietz 2008). In contrast with resilience research, vulnerability research provides a strong critique of the technocratic focus of earlier geophysical approaches (Miller et al. 2010), putting issues of power, inequality, and deprivation center stage (Doorn 2017). Part of the critique of the recent "resilience turn" is that it may indeed involve a shift back to these technocratic approaches.

Relation between the system level and the individuals within that system: The third point of criticism follows from the previous point. What is the position of individuals if resilience is primarily about the functioning of a system as a whole? As noted above, if resilience research covers the same ground as vulnerability research, then talking about resilience would be merely a matter of rhetoric with little added value. However, emphasizing the systemic part of resilience without looking at what this entails for the individual within that system overlooks important normative aspects (Berkes and Ross 2013; Cote and Nightingale 2011). That is, the relationship between the individual and the system in resilience highlights the role of normativity, mentioned in Section 12.2: interpretations of resilience can differ in their normativity. Some interpretations may refer to resilience as purely instrumental to achieving some goal, but other interpretations may refer to resilience as desirable in itself. Moreover, the fact that resilience may lead to undesirable outcomes (e.g., unjust outcomes) does not make the concept itself non-normative (Cañizares, Copeland, and Doorn 2021). These latter issues are often overlooked even in the literature that highlights the normativity of resilience, which is problematic as we will show below.

Most criticisms should not be seen as intractable, but instead as pointers to issues that should be included in a comprehensive conceptualization of resilience. What the points raised in this section highlight are that resilience is inevitably a normative concept or application of a concept, and eliding this normativity means inappropriately conflating resilience and vulnerability and failing to attend to the relationship between the individual and the system that resilience frameworks imply. Ultimately, what is needed is a conceptualization of resilience that is able to combine both the systematic character of resilience and the social and normative aspects that are part of the community framing. This in turn opens up opportunities to assess both the resilience of a system as well as its value, evaluating both the status quo and the ideal resilient city.

12.4 A Conceptualization

In order to develop a normative notion of resilience that can account for responsibilities, let us explore its different elements by formalizing the concept of resilience based on how the term is used in different disciplines. It goes too far to discuss all the different definitions (for recent overviews, see Béné et al. 2017; Doorn, Gardoni, and Murphy 2019; Meerow and Stults 2016), so we begin instead with a general and often-cited taxonomy that is provided by Folke (2006), who distinguishes between three notions of resilience, ranging from a narrow interpretation[1] of resilience, to ecological/ecosystem and social resilience, to an even broader social-ecological interpretation of resilience. It is generally considered an emergent property, where the system can be considered resilient if the different components can jointly accommodate and recover from shocks and thereby contribute to retaining the functions of the system as a whole (Da Silva, Kernaghan, and Luque 2012; Walker et al. 2006). However, this emergent character is not part of the most basic definition of resilience, which is the ability of a system to maintain its functions after disturbance. Consequently, we begin with a basic concept, closely related to the way Holling described resilience in ecology, as a system's buffer capacity and ability to withstand shocks and maintain its functions. This could be written in the form of a formula as follows:

Resilience$_1$: = the ability of system S to maintain its functions F after disturbance D.

The elements in this formula and the schematic letters used for them are as follows:

S: the entity (system) to which the label resilience applies.
F: the functions that the system should be able to fulfill in order to count as resilient.

The second description of resilience provided by Folke (2006) adds the element of self-organization and, here, the emergent character becomes more important. While self-organization is a difficult concept to formalize, at a minimum it pre-supposes that there are elements within the system that somehow relate to the overall functioning of the system. In a social context, this would mean that the elements that constitute the system should be able to ensure that the system functions as it should. A richer formalization of resilience therefore reads as follows:

Resilience$_2$: = the ability of system S to maintain its functions F through the actions A of its components C after disturbance D.

The last description provided by Folke defines resilience as the ability to learn and adapt. Here it would be interesting to look at how resilience engineering replaces traditional approaches in risk management that focus on the prevention of failure (Doorn 2021). In this resilience engineering paradigm, a resilient system is a system that is able to show successful behavior in a changing environment, where this changing environment is not necessarily conceived of as one of threats, but rather one of change and surprises. In other words, it is not known what the threats are, how the environment will change, in what direction or at what speed.[2] One way to formalize this is by generalizing the "disturbance D" of Resilience$_2$ to the more open "changing situations." Also the preposition "after" suggests that resilience is limited to reactive recovery after some disturbance. Further, within this formulation, it is implied that the action denoted by A occurs after the disturbance, whereas actions toward resilience occur before and during a disturbance as well. A more general formulation that accommodates all kinds of relevant action, as well as allowing for learning and adaptation, would therefore read as follows:

Resilience$_3$: = the ability of system S to maintain its functions F through the actions A of its components C in changing situations Sc.

Let us now see how we can give substance to the different letters in Resilience$_3$. First, we can think of the system S to which the label applies in terms of, for example, a specific community. In the case of climate adaptation, the community will often be a localized one, for example, within a neighborhood or a city, but in relation to other threats, the community may be much more dispersed geographically (cf. terrorism, virtual threats, migration). The exact demarcation of a community is far from trivial. First, there may be a difference between the community in terms of geographical area and the community in terms of population. The responses to Hurricane Katrina, for example, highlighted these different ways of framing the New Orleans community, with differing actions entailed by each framing, where preserving geographical community boundaries entailed trading off resilience to flooding for resilience to community dissolution. Second, even if geographical location and population more or less overlap, the question of defining a system's boundary is a very relevant question from the point of view of justice. Drawing the system's boundaries inevitably prompts the distributive question of who is entitled to membership in the community of justice (Dobson 1998). Thus, it seems we cannot give content to S in our formula until we have considered the components of that system, determined which are its key functions to preserve, as well as pinpointed what changing situations will require adaptation.

The crucial first step, then, seems to be to give substance to the functions F and the components C. These variables give answers to the question, "resilience of what." Let us start with the components C. We suggest taking humans as the primary components of the system. True, the people who together constitute the community may need resources and infrastructure, but people are the components of the system who act. Other components of these socio-technical systems can be considered supporting resources that enable or constrain the possible actions that people can take.[3]

The functions F that a community should be able to fulfill is again clearly a normative question. Whereas these functions are used descriptively in the biology and ecology literature and may evolve over time without normative repercussions, at the most basic level, the functions that a community fulfills vis-à-vis its members are inextricably linked to the question of what a good society is. Candidates for functions here are providing a safe, secure, and/or livable place for humans to live in. What the exact function is is context-dependent, but it should probably at least provide a place where people's basic rights are respected. In the scarce literature on resilience ethics specifically devoted to issues of justice, the capability approach has been suggested as a normative theory to give substance to the "functions" that a society should be able to fulfill (Doorn 2019).

Sc stands for the changing situations and is the answer to the question "resilience to what." Named after the publication by Carpenter et al. (2001), a common question formulated in the context of resilience is "resilience from what against what", but the very idea of resilience is exactly that it is often not known what the second "what" (the threats) are. The term "changing situations" allows for some specification without the need to define what the exact changes are. However, if resilience is to make an impact on policy making, it is of course necessary to provide the relevant context—for example, whether resilience is discussed in the context of, say, climate change or an aging society. Thus, we need to answer this question clearly each time we engage resilience as a guide to making policy; when we specify the changing situations that we want to respond to or prepare for, we have to identify the corresponding functions and components that also determine what actions will be needed to enact the policy that guides response and preparation.

The most difficult part of the formula to translate to the actual context is probably the actions A. Do the actions refer to incidental acts performed by people that happen to be successful or to specific tasks or obligations? To conceive of these actions as specific, maybe even pre-defined obligations or responsibilities, seems to go against the emergent character of resilience; that is, they can only be defined along with determining the content of F and C, who will act and toward the preservation of which functions. However, in the social science literature, resilience is often said to involve an implicit and unacknowledged transfer of responsibility from government toward citizens and other private actors (Hegger et al. 2017). The formula is designed to make these responsibilities more explicit. This in turn prompts the question of how closely related our normative notion of community resilience stays to the original idea of ecosystem resilience?

One way to address this question is by distinguishing between two paradigmatic situations: one where the behavior of the system can be characterized by causality and one where the behavior of the system can be characterized by emergence. In the case of causality, the actions A refer to specific responsibilities that are relatively easy to assign based on some desired outcome. For example, there seems to be a more or less direct causal relation between the amount of unpaved surfaces in a city and the drainage system on the one hand and the occurrence of so-called water nuisance (flooded streets) on the other. Here, citizens could, in principle, be given a responsibility to reduce the size of paved surface in their garden. True, it is the cumulative effect of many paved gardens that will lead to water nuisance (with the risk that a problem of many hands [Van de Poel et al. 2012] occurs), but, in principle, the relation between a citizen acting in a particular way (paving or not paving one's garden) and the outcome (nuisance) is a direct one. If we knew how all actors would act, then we could, in principle, predict the outcome of these joint actions. In those situations, it does seem to make sense to talk about task-responsibilities, and in these cases, the content of F and C will provide the content of A fairly directly.

At the other side of the spectrum, however, we have situations that are fully characterized by emergence. Here, we cannot trace the outcome causally back to the action performed by single individuals. Rather, any changes in Sc will result in changes to F, C, and A: depending on the changing situation, different actors and different functions may become more or less relevant and different actions possible. It is the interaction between different people acting and interacting that leads to some outcome that cannot be predicted. The S in these cases is emergent from the variables, which change along with each other. In such cases, S is emergent and changes along with the situation itself, and for such systems, it seems problematic to assign specific responsibilities to individual citizens. An example of such emergent behavior is the situation after a natural hazard, for example, flooding. In a case of flooding, mass evacuation often leads to traffic queues that ultimately lead to fewer people being able to escape the dangerous situation. So, here, the "ideal" behavior is probably to be compared with a swarm of sparrows that adapt to the situation. This means that some people make use of "vertical evacuation" (that is, flying to high-rise buildings with sufficient resources to survive some days) and some of "horizontal evacuation" (flying to locations that are not flooded). Additionally, in a flood event, people should not only keep themselves safe, but at least a sufficiently large proportion of people should also be available to support the more vulnerable people. Which citizens should opt for vertical evacuation and which for horizontal evacuation is impossible to predict beforehand as it may depend on contextual factors and it may change over time. In such a situation, assigning citizens a very strict responsibility seems difficult as it is not known beforehand what exactly is needed from each of the citizens. What is needed are the right conditions for this emergent behavior to hold. Instead of talking about citizen responsibilities, it may therefore make more sense to talk of the government or some other public actor as being responsible for creating the conditions so that the acts of the different components are most likely to lead to the desired outcomes. In other words, resilience policy should go hand in hand with active involvement from governments to create the conditions for the desired emergent effect.

Applying the formula that we have proposed, therefore, requires not only answering the questions, resilient of what and to what, but also addressing the relationships between the elements that constitute the relevant system. That is, S cannot be determined until the other elements in the formula have been identified. Therefore, the process of deliberation required to give content to the variables requires *explicit* attention to the normative processes involved in identifying which components have to act in order to preserve which functions in what context of contingency and change. It is the relationships between the elements within the system that constrain and enable actions and, thereby, determine what policy and actions toward resilience of the system at hand are possible. Consequently, responsibilities are part and parcel of the system itself: in defining the system S according to a composition of F, C, and A together, responsibility arrangements become part of the description of that system and thereby an outcome of the process of deliberation itself.

In turn, the formula we have suggested here provides guidance as to when decisions about policy and prescriptions for action need to be reconsidered: that is, whenever one of the variables is altered, the description of the system itself – and the responsibility arrangements included within that description – needs to be reconfigured.

While our formula does not resolve the critiques above, by eliminating the conflicts identified in the critiques, it does allow us to avoid their problematic implications. For instance, the distinction we draw between causal and emergent situations demonstrates how the formula allows us to build a bridge between individual actions and the normative ideal that shapes the system. By integrating C into the formula as a variable that has a direct impact on S, we provide space for deliberating about that relationship between the individual and the system through addressing how A fits in with C and F in an explicit way. Rather than a technocratic resolution for vulnerabilities, the formula gives us the means to engage in ongoing deliberation about the context of vulnerabilities and the influence of the system on these vulnerabilities without conflating vulnerabilities and resilience. In each instance of using the formula, we call into question the meaning of resilience in relation to the status quo, desirable functions, and components, and, therefore, we avoid slipping into common tropes of resilience without questioning them.

12.5 Concluding Remarks

In this chapter, we developed a formula that allows for a more transparently normative analysis of community resilience. The elements in the formula are not intended as a blueprint for what it is to make something resilient. Rather, they should be seen as elements to consider when assessing the resilience of a community while also attending to issues of normativity, justice, and responsibility arrangements as they emerge in adaptive contexts. We distinguished between two extreme situations. At one end of the extreme, there is a direct causal relation between the actions of the actors in the system and the behavior at the system level. Here, responsibilities are relatively easy to assign. At the other extreme, the behavior of the system is characterized by emergence and it cannot directly be traced back to the behavior of the individual actors in the system. Here, it seems difficult to assign specific responsibilities to individual actors. Instead, it makes more sense to create the conditions that make the desired emergent behavior more likely. This suggests that the use of the term resilience should maybe not be taken too literally, but rather be seen as a metaphor for how society can deal with changing situations.

Notes

[1] In most ecological and social science literature on resilience, this narrow interpretation is often referred to as "engineering resilience." We think this is misleading, as this narrow resilience is not the same as the definition of resilience

that is common in the field of engineering. It is also ambiguous in the sense that the term "resilience engineering" refers to a specific approach within safety engineering for dealing with risks. This resilience engineering approach goes exactly against the narrow view that is also being criticized by Holling as not being applicable in the contexts of ecosystems.

[2] In fact, transformative approaches to resilience as "bounce-forward" suggest that keeping this open is a necessary implication of resilience as a strategy for changing circumstances. That is, if we interpret all potential changes as threats, then we have assumed that the status quo is sufficiently ideal to preserve (and that transformation is not a suitable goal) and thus failed to motivate support for transformative efforts or other improvements via resilience-based planning. Rather, any potential threat, under an adaptive model, also has the potential to be merely a changing circumstance if, for example, the result is positive for the system and the individuals within it.

[3] Note that our focus here on urban resilience allows us to focus on human actors as well; in other contexts, non-human actors may also fit into the responsibility model, but we leave that for further exploration beyond the bounds of this chapter.

References

4TU.Resilience Engineering. 2021. "DeSIRE Mission Statement." Accessed 19 October.

Adger, W. Neil. 2000. "Social and Ecological Resilience: Are They Related?" *Progress in Human Geography* 24 (3): 347–64. http://dx.doi.org/10.1191/030913200701540465.

Adger, W. Neil, Katrina Brown, Donald R. Nelson, Fikret Berkes, Hallie Eakin, Carl Folke, Kathleen Galvin, Lance Gunderson, Marisa Goulden, Karen O'Brien, Jack Ruitenbeek, and Emma L. Tompkins. 2011. "Resilience Implications of Policy Responses to Climate Change." *Wiley Interdisciplinary Reviews: Climate Change* 2 (5): 757–66. http://dx.doi.org/10.1002/wcc.133.

Adger, W. Neil, Suraje Dessai, Marisa Goulden, Mike Hulme, Irene Lorenzoni, Donald R. Nelson, Lars Otto Naess, Johanna Wolf, and Anita Wreford. 2009. "Are There Social Limits to Adaptation to Climate Change?" *Climate Change* 93: 335–54.

Barnett, Jon. 2001. "Adapting to Climate Change in Pacific Island Countries: The Problem of Uncertainty." *World Development* 29 (6): 977–93. http://dx.doi.org/10.1016/s0305-750x(01)00022-5.

Béné, Christophe, Lyla Mehta, Gordon McGranahan, Terry Cannon, Jaideep Gupte, and Thomas Tanner. 2017. "Resilience As a Policy Narrative: Potentials and Limits in the Context of Urban Planning." *Climate and Development*: 1–18. http://dx.doi.org/10.1080/17565529.2017.1301868.

Berkes, Fikret. 2017. "Environmental Governance for the Anthropocene? Social-Ecological Systems, Resilience, and Collaborative Learning." *Sustainability* 9 (7): 1232.

Berkes, Fikret, and Helen Ross. 2013. "Community Resilience: Toward an Integrated Approach." *Society & Natural Resources* 26 (1): 5–20. http://dx.doi.org/10.1080/08941920.2012.736605.

Biggs, Reinette, Maja Schlüter, and Michael L. Schoon (Eds.). 2015. *Building Principles for Resilience: Sustaining Ecosystem Services in Social-Ecological Systems*. Cambridge, UK: Cambridge University Press.

Brown, Katrina, and Elizabeth Westaway. 2011. "Agency, Capacity, and Resilience to Environmental Change: Lessons from Human Development, Well-Being, and Disasters." *Annual Review of Environment and Resources* 36 (1): 321–42. http://dx.doi.org/10.1146/annurev-environ-052610-092905.

Butler, Catherine, and Nicolas Frank Pidgeon. 2011. "From 'Flood Defence' to 'Flood Risk Management': Exploring Governance, Responsibility, and Blame." *Environment and Planning C-Government and Policy* 29 (3): 533–47. http://dx.doi.org/10.1068/c09181j.

Cañizares, Jose C., Samantha M. Copeland, and Neelke Doorn. 2021. "Making Sense of Resilience." *Sustainability* 13 (15): 8538. http://dx.doi.org/10.3390/su13158538.

Cannon, Terry, and Detlef Müller-Mahn. 2010. "Vulnerability, Resilience and Development Discourses in Context of Climate Change." *Natural Hazards* 55 (3): 621–35. http://dx.doi.org/10.1007/s11069-010-9499-4.

Carpenter, Steve, Brian Walker, J. Marty Anderies, and Nick Abel. 2001. "From Metaphor to Measurement: Resilience of What to What." *Ecosystems* 4 (8): 765–81.

Chandra, Anita, Joie Acosta, Stefanie Stern, Lori Uscher-Pines, Malcolm V. Williams, Douglas Yeung, Jeffrey Garnett, and Lisa S. Meredith. 2010. *Building Community Resilience to Disasters: A Way Forward to Enhance National Health Security*. Santa Monica: RAND Corporation.

Connor, Katherine M., and Jonathan R.T. Davidson. 2003. "Development of a New Resilience Scale: The Connor-Davidson Resilience Scale (CD-RISC)." *Depression and Anxiety* 18 (2): 76–82.

Copeland, Samantha, Tina Comes, Sylvia Bach, Michael Nagenborg, Yannic Schulte, and Neelke Doorn. 2020. "Measuring Social Resilience: Trade-Offs, Challenges and Opportunities for Indicator Models in Transforming Societies." *International Journal of Disaster Risk Reduction* 51: 101799. https://doi.org/10.1016/j.ijdrr.2020.101799.

Cote, Muriel, and Andrea J. Nightingale. 2011. "Resilience Thinking Meets Social Theory." *Progress in Human Geography* 36 (4): 475–89. http://dx.doi.org/10.1177/0309132511425708.

Cretney, Raven. 2014. "Resilience for Whom? Emerging Critical Geographies of Socio-Ecological Resilience." *Geography Compass* 8 (9): 627–40. http://dx.doi.org/10.1111/gec3.12154.

Da Silva, Jo, Sam Kernaghan, and Andres Luque. 2012. "A Systems Approach to Meeting the Challenges of Climate Change." *International Journal of Urban Sustainable Development* 4 (2): 125–45.

Davoudi, Simin. 2012. "Resilience: A Bridging Concept or a Dead End?" *Planning Theory & Practice* 13 (2): 299–307. http://dx.doi.org/10.1080/14649357.2012.677124.

Deppisch, Sonja. 2017. "Cities and Urban Regions Under Change: Between Vulnerability, Resilience, Transition and Transformation." In *Urban Regions Now & Tomorrow: Between Vulnerability, Resilience and Transformation*, edited by Sonja Deppisch, 1–16. Dordrecht: Springer.

Dobson, Andrew. 1998. *Justice and the Environment: Conceptions of Environmental Sustainability and Dimensions of Social Justice*. Oxford: Oxford University Press.

Doorn, Neelke. 2017. "Resilience Indicators: Opportunities for Including Distributive Justice Concerns in Disaster Management." *Journal of Risk Research* 20 (6): 711–31. http://dx.doi.org/10.1080/13669877.2015.1100662.

Doorn, Neelke. 2019. "How Can Resilient Infrastructures Contribute to Social Justice? – Preface to the Special Issue of Sustainable and Resilient Infrastructure on Resilient Infrastructures and Social Justice." *Sustainable and Resilient Infrastructure* 4 (3): 99–102. http://dx.doi.org/10.1080/23789689.2019.1574515.

Doorn, Neelke. 2021. "The Role of Resilience in Engineering." In *Handbook of Philosophy of Engineering*, edited by Diane P. Michelfelder and Neelke Doorn, 482–93. New York: Routledge.

Doorn, Neelke, Lieke Brackel, and Sara Vermeulen. 2021. "Distributing Responsibilities for Climate Adaptation: Examples from the Water Domain." *Sustainability* 13 (7): 3676. http://dx.doi.org/10.3390/su13073676.

Doorn, Neelke, Paolo Gardoni, and Colleen Murphy. 2019. "A Multidisciplinary Definition and Evaluation of Resilience: The Role of Social Justice in Defining Resilience." *Sustainable and Resilient Infrastructure* 4 (3): 112–23. http://dx.doi.org/10.1080/23789689.2018.1428162.

Eraydin, Ayda, and Tuna Taşan-Kok (Eds.). 2013. *Resilience Thinking in Urban Planning*. Dordrecht: Springer.

Folke, Carl. 2006. "Resilience: The Emergence of a Perspective for Social-Ecological Systems Analyses." *Global Environmental Change* 16 (3): 253–67.

Folke, Carl, Stephen R. Carpenter, Brian Walker, Marten Scheffer, Thomas Elmqvist, Lance Gunderson, and Crawford S. Holling. 2004. "Regime Shifts, Resilience, and Biodiversity in Ecosystem Management." *Annual Review of Ecology Evolution and Systematics* 35: 557–81. http://dx.doi.org/10.1146/annurev.ecolsys.35.021103.105711.

Folke, Carl, Stephen R. Carpenter, Brian Walker, Marten Scheffer, F. Stuart Chapin, and Johan Rockstrom. 2010. "Resilience Thinking: Integrating Resilience, Adaptability and Transformability." *Ecology and Society* 15 (4): 20–8.

Folke, Carl, Åsa Jansson, Johan Rockström, Per Olsson, Stephen R. Carpenter, F. Stuart Chapin, Anne-Sophie Crépin, Gretchen Daily, Kjell Danell, Jonas Ebbesson, Thomas Elmqvist, Victor Galaz, Fredrik Moberg, Måns Nilsson, Henrik Österblom, Elinor Ostrom, Åsa Persson, Garry Peterson, Stephen Polasky, Will Steffen, Brian Walker, and Frances Westley. 2011. "Reconnecting to the Biosphere." *AMBIO: A Journal of the Human Environment* 40 (7): 719–38. http://dx.doi.org/10.1007/s13280-011-0184-y.

Fünfgeld, Hartmut, and Darryn McEvoy. 2012. "Resilience As a Useful Concept for Climate Change Adaptation?" *Planning Theory & Practice* 13 (2): 324–8. http://dx.doi.org/10.1080/14649357.2012.677124.

Gallopín, Gilberto C. 2006. "Linkages between Vulnerability, Resilience, and Adaptive Capacity." *Global Environmental Change* 16 (3): 293–303.

Gunderson, Lance. 2010. "Ecological and Human Community Resilience in Response to Natural Disasters." *Ecology and Society* 15 (2): 1–11.

Hamel, Gary, and Liisa Valikangas. 2003. "The Quest for Resilience." *Harvard Business Review* 81 (9): 52–65.

Hegger, Dries L. T., Heleen L. P. Mees, Peter P. J. Driessen, and Hens A. C. Runhaar. 2017. "The Roles of Residents in Climate Adaptation: A Systematic Review in the Case of the Netherlands." *Environmental Policy and Governance* 27 (4): 336–50. http://dx.doi.org/10.1002/eet.1766.

Holling, Crawford S. 1973. "Resilience and Stability of Ecological Systems." *Annual Review of Ecology and Systematics* 4: 1–23.

IPCC. 2022. *Climate Change 2022: Impacts, Adaptation, and Vulnerability*. Cambridge, United Kingdom and New York: Cambridge University Press.

Janssen, Marco A., Michael L. Schoon, Weimao Ke, and Katy Börner. 2006. "Scholarly Networks on Resilience, Vulnerability and Adaptation within the Human Dimensions of Global Environmental Change." *Global Environmental Change* 16 (3): 240–52. https://doi.org/10.1016/j.gloenvcha.2006.04.001.

Jordan, Elizabeth, and Amy Javernick-Will. 2013. "Indicators of Community Recovery: Content Analysis and Delphi Approach." *Natural Hazards Review* 14 (1): 21–8. http://dx.doi.org/10.1061/(ASCE)NH.1527-6996.0000087.

Leichenko, Robin. 2011. "Climate Change and Urban Resilience." *Current Opinion in Environmental Sustainability* 3 (3): 164–8. https://doi.org/10.1016/j.cosust.2010.12.014.

Martin, Ron, and Peter Sunley. 2015. "On the Notion of Regional Economic Resilience: Conceptualization and Explanation." *Journal of Economic Geography* 15 (1): 1–42. http://dx.doi.org/10.1093/jeg/lbu015.

McAslan, Alastair. 2010. *The Concept of Resilience. Understanding Its Origins, Meaning and Utility*. Adelaide: The Torrens Resilience Institute.

McEvoy, Darryn, Hartmut Fünfgeld, and Karyn Bosomworth. 2013. "Resilience and Climate Change Adaptation: The Importance of Framing." *Planning Practice & Research* 28 (3): 280–93. http://dx.doi.org/10.1080/02697459.2013.787710.

McLaughlin, Paul, and Thomas Dietz. 2008. "Structure, Agency and Environment: Toward an Integrated Perspective on Vulnerability." *Global Environmental Change* 18 (1): 99–111. https://doi.org/10.1016/j.gloenvcha.2007.05.003.

Meerow, Sara, Joshua P. Newell, and Melissa Stults. 2016. "Defining Urban Resilience: A Review." *Landscape and Urban Planning* 147: 38–49.

Meerow, Sara, and Melissa Stults. 2016. "Comparing Conceptualizations of Urban Climate Resilience in Theory and Practice." *Sustainability* 8 (7): 701.

Miller, Fiona, Henny Osbahr, Emily Boyd, Frank Thomalla, Sukaina Bharwani, Gina Ziervogel, Brian Walker, Joern Birkmann, Sander van der Leeuw, Johan Rockstroem, Jochen Hinkel, Tom Downing, Carl Folke, and Donald Nelson. 2010. "Resilience and Vulnerability: Complementary or Conflicting Concepts?" *Ecology and Society* 15 (3): 1–25.

Norris, Fran H., Susan P. Stevens, Betty Pfefferbauam, Karen F. Wyche, and Rose L. Pfefferbaum. 2008. "Community Resilience As a Metaphor, Theory, Set of Capacities, and Strategy for Disaster Readiness." *American Journal of Community Psychology* 41 (1–2): 127–50.

Paton, Douglas, and David Johnston. 2001. "Disasters and Communities: Vulnerability, Resilience and Preparedness." *Disaster Prevention and Management* 10 (4): 270–7.

Pendall, Rolf, Kathryn A. Foster, and Margaret Cowell. 2010. "Resilience and Regions: Building Understanding of the Metaphor." *Cambridge Journal of Regions, Economy and Society* 3 (1): 71–84. http://dx.doi.org/10.1093/cjres/rsp028.

Porter, Libby, and Simin Davoudi. 2012. "The Politics of Resilience for Planning: A Cautionary Note." *Planning Theory & Practice* 13 (2): 329–33. http://dx.doi.org/10.1080/14649357.2012.677124.

Sharifi, Ayyoob, and Yoshiki Yamagata. 2016. "Principles and Criteria for Assessing Urban Energy Resilience: A Literature Review." *Renewable and Sustainable Energy Reviews* 60: 1654–77. http://dx.doi.org/10.1016/j.rser.2016.03.028.

Shaw, Keith, and Louise Maythorne. 2013. "Managing for Local Resilience: Towards a Strategic Approach." *Public Policy and Administration* 28 (1): 43–65. http://dx.doi.org/10.1177/0952076711432578.

Shutters, Shade T., Rachata Muneepeerakul, and José Lobo. 2015. "Quantifying Urban Economic Resilience Through Labour Force Interdependence." *Palgrave Communications* 1 (15010): 1–7. http://dx.doi.org/10.1057/palcomms.2015.10.

Southwick, Steven M., Meena Vythilingam, and Dennis S. Charney. 2005. "The Psychobiology of Depression and Resilience to Stress: Implications for Prevention and Treatment." *Annual Review of Clinical Psychology* 1: 255–91. http://dx.doi.org/10.1146/annurev.clinpsy.1.102803.143948.

Turner, Billie L. 2010. "Vulnerability and Resilience: Coalescing or Paralleling Approaches for Sustainability Science?" *Global Environmental Change* 20 (4): 570–6. http://dx.doi.org/10.1016/j.gloenvcha.2010.07.003.

Twigger-Ross, Clare, Terry Coates, Hugh Deeming, Paula Orr, Mark Ramsden, and John Stafford. 2011. *Community Resilience Research: Final Report on Theoretical Research and Analysis of Case Studies Report to the Cabinet Office and Defence Science and Technology Laboratory*. London: Collingwood Environmental Planning Ltd.

Twigger-Ross, Clare, Kashefi Elham, Sue Wheldon, Katya Brooks, Hugh Deeming, Steven Forrest, Jane Fielding, Alan Gomersall, Tim Harries, Simon McCarthy, Paula Orr, Dennis Parker, Sue Tapsell. 2014. *Flood Resilience Community Pathfinder Evaluation: Rapid Evidence Assessment*. London: Defra.

Van de Poel, Ibo R., Jessica A. Nihlén Fahlquist, Neelke Doorn, Sjoerd J. Zwart, and Lamber M.M. Royakkers. 2012. "The Problem of Many Hands: Climate Change As an Example." *Science and Engineering Ethics* 18 (1): 49–67.

Walker, Brian, Lance Gunderson, Ann Kinzig, Carl Folke, Steve Carpenter, and Lissen Schultz. 2006. "A Handful of Heuristics and Some Propositions for Understanding Resilience in Social-Ecological Systems." *Ecology and Society* 11 (1): 1–13.

Walker, Brian, C.S. Holling, Stephen R. Carpenter, and Ann Kinzig. 2004. "Resilience, Adaptability and Transformability in Social-Ecological Systems." *Ecology and Society* 9 (2): 5–13.

Wardekker, A. 2022. "Framing 'Resilient Cities': System Versus Community Focussed Interpretations of Urban Climate Resilience." In *Urban Resilience: Methodologies, Tools and Evaluation*, edited by Luis Jiménez Herrero, Octavio González Castillo, Josep Pont Vidal and Ernesto Santibanez. Berlin: Springer.

Wardekker, J. Arjan, Arie De Jong, Joost M. Knoop, and Jeroen P. Van der Sluijs. 2010. "Operationalising a Resilience Approach to Adapting an Urban Delta to

Uncertain Climate Changes." *Technological Forecasting & Social Change* 77 (6): 987–98.

Woods, David D., and John Wreathall. 2003. *Managing Risk Proactively: The Emergence of Resilience Engineering*. Columbus: Institute for Ergonomics, The Ohio State University.

Zurlini, Giovanni, Irene Petrosillo, K. Bruce Jones, and Nicola Zaccarelli. 2013. "Highlighting Order and Disorder in Social–Ecological Landscapes to Foster Adaptive Capacity and Sustainability." *Landscape Ecology* 28 (6): 1161–73. http://dx.doi.org/10.1007/s10980-012-9763-y.

13 Individual Climate Risks at the Bounds of Rationality

Avram Hiller

Dès qu'il y a vie, il y a danger.

Madame de Staël (1845, 564)

13.1 Introduction

All ordinary decisions involve *some* risk. If I go outside for a walk, I may trip and injure myself. But if I don't go for a walk, I slightly increase my chances of cardiovascular disease. Typically, we disregard most small risks. When, for practical purposes, is it appropriate for one to ignore risk? This issue looms large because many activities performed by those in wealthy societies, such as driving a car, in some way risk contributing to climate harms. Are these activities morally appropriate?

In this chapter, I will argue that it is appropriate to ignore many small risks. I am not the first to argue for this conclusion. However, the reasons that I give for ignoring small risks differ to some extent from those identified by others in some recent debates. In particular, I will argue that because our rationality is *bounded*, it is impossible for us to include every small risk in our decision-making process, and so we may reasonably use heuristics to guide many decisions. Although our use of heuristics allows for the reasonable ignoring of some risks and perhaps explains why one might be inclined to think that individual climate-related risks are negligible, the main aim of this paper is to show that even reasonable use of heuristics does not permit the *general* ignoring of climate change-related risk by individuals.

The other main aim of this paper is expository. Philosophers have engaged with issues related to the present one in great detail, especially since Derek Parfit's *Reasons and Persons* (1984). Although in this paper I advance a particular thesis about when it is appropriate for individuals to ignore their greenhouse gas (hereafter GHG) emissions' climate-related risks, the relevant literature is vast and multidisciplinary (including climate science, environmental economics, behavioral economics/psychology, decision theory, ethics, and metaphysics). Because the evidence required to

DOI: 10.4324/9781003276029-18

assess individual climate-relevant obligations is so large and disparate, I have attempted to bring together several literatures that pertain to the question. Although I am not able to be comprehensive in my discussion of the relevant literatures, I hope that readers will benefit from my efforts to bring together considerations from several different fields even if they disagree with my conclusions.

I should note that economists and decision theorists sometimes distinguish between *risk* and *uncertainty*. The former is when there are specifiable probabilities of an outcome occurring, and the latter – *Knightian uncertainty* (from Knight 1921) – is when there are not. In this paper, I will not distinguish between the two and will use the notion of *risk* to cover cases when there is some possibility (whether it be precisely known, imprecise, or unspecifiable) of some negatively valenced outcome occurring. Cases of individual climate risk are ones of Knightian uncertainty.

13.2 "It Makes No Difference," *Again*?

A number of philosophers (Sinnott-Armstrong 2005, and later Kingston and Sinnott-Armstrong 2018, as well as Cripps 2013; Gesang 2017; Nefsky 2012; Sandberg 2011) argue that, despite the real existence of global climate change and its harms, individual actions do not make a difference in regard to climate change. The general idea, in Kingston and Sinnott-Armstrong (2018), is that the effects of one individual emitting GHGs are so small and diffuse in the global causal structure of climate change that it makes no difference whether one individual does or does not emit GHGs at that scale.[1]

In this paper, I will not engage directly with these arguments to any significant extent. My own view, mostly in line with Broome (2019, 2021; also see Hiller 2011a, 2011b; Morgan-Knapp and Goodman 2015; Nye 2021), is that (A) GHG-emitting actions should be held to have some non-negligible amount of *expected* negative disvalue and that (B) this expected negative disvalue is morally relevant. Roughly speaking, Broome (2019) shows that individual GHG-emitting activities have a strictly increasing *expected* harm function even if they do not in fact cause harm or trigger a harm threshold. This is not to say that actions that emit GHGs are never all-things-considered permissible; to determine whether they are permissible depends upon one's overall normative views and other relevant facts. The maximizing consequentialist, for instance, will hold that a GHG-emitting action is permissible if it is the best action one might take. Perhaps the expected *benefits* from certain GHG-emitting actions are high, or perhaps one has no other *better* option than to emit GHGs. But the debate surrounding individual climate ethics following Sinnott-Armstrong (2005) has largely been occupied with the question of whether there should be *anything* on the negative side of the ledger from the expected climate harms due to individual actions, and it is this question that I take to be answered in the affirmative.

Although I will raise some concerns about some details of Broome's analysis, in this paper, I will primarily take a broader look at the debate. Is standard expected value theory applicable to individual climate ethics in the first place? At the very least, the application of expected value theory to climate change is *fraught*: in one formulation, it involves taking a tiny fraction – the proportion of the total amount of global GHG emissions that one individual is responsible for – and multiplying it by a huge value – the total amount of harm that can be expected to occur (in the form of a harm function) given various climate scenarios. While Nolt (2011), Broome (2019), and Hiller (2011a, 2011b) unapologetically perform this multiplication and hold that the resultant product is a non-negligible value, the very consideration of a tiny risk of harm may be objectionable for independent reasons.

Within the field of risk analysis, for example, some have argued that certain small risks, referred to as *de minimis* risks, may be reasonably ignored. Peterson (2002) gives the first extended philosophical analysis and critique of *de minimis* risk (also see Adler 2007; Lundgren and Orri Stefánsson 2020 for related criticisms). The idea behind *de minimis* risk is that some risks are too small to merit consideration by the law. For instance, the law need not require that all carcinogenic items be omitted from food items, because potentially all items may raise cancer risk by the tiniest of amounts. As Peterson notes (2002, 48), initially the notion of *de minimis* risk was intended as a response to an extreme form of the *precautionary principle*, since a *de minimis* risk principle says that precautions need not be undertaken when risks are extremely small. Still, the notion of *de minimis risk* can also be used in the context of a less extreme form of the precautionary principle, because when there is a tiny risk of a potentially large catastrophe, policymakers may still wish to ignore that possibility. Even Cass Sunstein, perhaps the most ardent advocate of cost-benefit analysis, supports the employment of a *de minimis* risk principle (2002, 193–5; 214–6).

As Peterson argues, the notion of a *de minimis* risk admits vagueness – there is no sharp boundary between the *de minimis* cases and the non-*de minimis* cases. Lundgren and Orri Stefánsson (2020, 913) further argue that in cases where there is a tiny risk to one option, and no benefit, then it would be wrong to impose the risk. These arguments seem to me to be correct. The literature on *de minimis* risk has largely pertained to risks from governmental policies, and I will bracket it for the moment.

A related literature has arisen recently regarding the notion of *discounting* small probabilities.[2] Kosonen (2021), Monton (2019), and Smith (2014) have argued that we may rationally discount tiny probabilities. For instance, Bostrom's (2009) case of Pascal's Mugging seems to show that without discounting small probabilities, expected value theory is susceptible to a seemingly absurd result of indicating that it is best to sacrifice finite goods for an astronomically unlikely promise of enormous rewards. Balfour (2021) explicitly connects this argument to existential risk for humans, arguing that we must thus make extreme and ridiculous efforts to reduce existential risk, and suggests that expected value theory should be abandoned for this reason.

In response, a number of commentators have responded that discounting itself leads to absurdity. For instance, both Ebert et al. (2020, 438–9) and Barrington (ms.) argue that partitioning the outcome space of a choice into multiple partitions may turn an intuitively wrong choice into an appropriate one if discounting of tiny risks occurs. If two possible bad outcomes each independently have tiny risks, then these can be ignored according to a discounting principle, but if the outcome were simply a single outcome with the sum of these risks, then it could not be properly ignored. But this ramification of discounting seems inappropriate.

Gesang (2021) argues that individual actions are subject to inefficacy and inscrutability concerns and thus are not subject to expected value analysis. At the same time, according to Gesang, *large* ones *do* make a significant expected difference. However, for reasons related to those just discussed, this argumentative move must be mistaken. For instance, Hiller (2011a, 2011b, 355) notes that drives themselves can be partitioned into smaller concatenated actions. If there is a threshold below which risks are morally negligible, then one could avoid having responsibility for many of one's culpable doings simply by dividing them into smaller ones. If going on a drive across the United States is above the culpability threshold for climate-based emissions risk, but driving shorter distances is not, one could simply plan to leave from one coast, drive from town to nearby town, eventually happily arriving at the other coast without the burden of a guilty conscience. But this seems absurd. Instead, traditional expected value theory seems vindicated, because it simply adds up the disvalue of the smaller drives into the same sum as the single longer one. *Causes* in group or large event phenomena, on the other hand, can't be decomposed into equal partitions. What my particular emissions will in fact trigger may be different from what your particular emissions will trigger. But *expected value* of large actions can be divided into partitions when there are no known reasons for making an exception for a particular marginal contribution.

Additionally, Barrington argues (ms., §3), similarly to Lundgren and Stefánsson, that when there are no *other* relevant factors, it would be wrong to ignore a tiny probability of harm simply because it is small. Indeed, it does not seem right that the mere *minimal* nature of some risks could, entirely on its own, be sufficient reason to ignore them.

Another longstanding argument in the literature has been more broadly viewed as successfully undermining traditional expected value theory – the *Small Improvements Argument*. Although the relationship between the Small Improvements Argument to individual climate ethics might seem at first glance distant,[3] what I will argue in the next two sections is that lessons from the psychology of decision-making that are revealed by a close look at the Small Improvements Argument can help shed light on individual climate ethics.

13.3 The Small Improvements Argument

It is a natural reaction in certain cases to think that certain small differences in outcomes should make no difference in the choiceworthiness of options. Joseph Raz (1986), following Ronald de Sousa (1974), gives one such case. Imagine a situation where a person is deciding upon a career as a philosopher or a career as a lawyer. Assume that neither option is better than the other with respect to success or desirability of career for the person, and the person finds it extremely difficult to choose. Plausibly enough, argues Raz, if the person learns that the legal career has a very slightly better salary than the person had previously considered, it is *still* the case that neither career choice is better than the other for the individual.

Several philosophers have concluded from this example that values are *incommensurable*, a claim that deserves fuller treatment elsewhere. What is relevant here is that this example purportedly shows that foundational principles from classical expected value theory do not hold. In particular, here are two core principles: (Let V(X) mean the value of state of affairs X.)

Completeness: for any states A, B, exactly one of the following is the case:

$$\mathrm{V}(A) = \mathrm{V}(B); \mathrm{V}(A) > \mathrm{V}(B); \text{or } \mathrm{V}(A) < \mathrm{V}(B)$$

Transitivity: If $\mathrm{V}(A) = \mathrm{V}(B)$, and $\mathrm{V}(B) = \mathrm{V}(C)$, then $\mathrm{V}(A) = \mathrm{V}(C)$

In Raz's case, let P denote the individual's career as a philosopher; let L denote the individual's career as a lawyer; Let $L+$ denote a career as a lawyer, but with it being slightly more lucrative than in L. The intuitive sets of claims are that

1. $\mathrm{V}(L) = \mathrm{V}(P)$
2. $\mathrm{V}(L+) = \mathrm{V}(P)$
3. $\mathrm{V}(L) < \mathrm{V}(L+)$

but by (1) and (2) and transitivity,

4. $\mathrm{V}(L) = \mathrm{V}(L+)$, violating completeness

Instead, Ruth Chang (2002) argues that the proper characterization of the situation is to say that the values of L and P are *on a par* (see Andreou 2015 for a more recent defense of the notion of parity).

Interestingly, one way to initially understand the Kingston/Sinnott-Armstrong view is to claim that if *D* is a typical Sunday drive, and *D*+ is a Sunday drive in a vehicle that emits no GHGs, that V(*D*) = V(*D*+). But, intuitively at least, V(*D*) < V(*D*+), perhaps violating completeness. Although one shouldn't say that *D* and *D*+ are *on a par*, Kingston and Sinnott-Armstrong would likely hold that for any choice when one is faced with a decision between *D* and some *other* option *O*, whether to choose *D* or *O* need not involve consideration of *D*'s GHG emissions, and thus the relevant features of the individual's deliberation should be the same as in deliberation between *D*+ and *O*. The choiceworthiness of going for a drive relative to some other option should be the same whether or not the drive emits GHGs.

One early response to the Small Improvements Argument is that it is an *epistemic* issue (Regan 1989, 1059–61). Regan's idea is that even if we intuitively think that both V(*L*) = V(*P*) and V(*L*+) = V(*P*), they are not both true, and this is because we are not properly grasping the relevant fine-grained states of affairs in question. We are simply uncertain of V(*L*), V(*L*+), and V(*P*) and are not really evaluating them.

I should mention that Ruth Chang responds (2002, 669–70) to the kind of concern raised by Regan by noting that the small improvement phenomenon occurs not just for major life decisions but also small decisions, like ones in which one decides whether to have a cup of coffee or tea. If one is undecided between the two, and then one hears that a slightly better tea is available, that will not necessarily sway one to choose the newly available tea over coffee. It is not my aim here to delve into all the details of the Small Improvements Argument, but I should note that this example has never seemed convincing to me. When one evaluates coffee against tea, it is not the coffee and tea that are intrinsically valuable; what matters are the experiences that they will produce in the drinker. But it is unclear how carefully the drinker can in fact anticipate the full set of experiences that they will have upon drinking coffee or tea, especially given that the value of aesthetically pleasing experiences is modulated by the context one is in. Even a choice between familiar coffee and familiar tea can have significant uncertainties in it in any new circumstance, and so the epistemic move still seems appropriate.

The general approach that I'd like to take in responding to the Small Improvements Argument is similar to that of Regan and of Anderson (2015). But I will express it using some notions from the field of behavioral economics.

13.4 Bounded Rationality and Minimal Risks

13.4.1 Bounded Rationality

The concept of *bounded rationality* is familiar to many (see Gigerenzer 2021 for a helpful history). Herbert Simon (1955) and later Amos Tversky and Daniel Kahneman (see 2011), and many others, argue that there are significant constraints on human abilities to reason. Some of these constraints can primarily be seen as endogenous to the human mind: humans can sometimes be slow in arriving at answers, and we are susceptible to biases that cause us, in certain contexts at least, to regularly provide incorrect answers. It is, famously, a controversial issue whether this should lead us to think of humans as being to a significant extent irrational or, rather, to understand rationality as necessarily contextual/ecological (as in Gigerenzer 2000). Other constraints are best seen as exogenous: we often have limited time and evidential resources to make judgments and decisions. Here, I will remain neutral on the question of the rationality of endogenous constraints on human judgment and decision-making and instead focus on constraints that are primarily exogenous.

Kahneman and Tversky, as well as Gigerenzer (2000), emphasize that people employ *heuristics* in making judgments and decisions. Although oftentimes heuristics are commonly thought of as *rules of thumb*, and Gigerenzer (2000) understands heuristics as tools, I'd like to employ a definition given by Kahneman in his later collaborative work with Shane Frederick (2002). As Kahneman (2003, 466) summarizes, "A judgment is said to be mediated by a heuristic when the individual assesses a specified *target attribute* of a judgment object by substituting a related *heuristic attribute* that comes more readily to mind." So rather than use cognitive resources in assessing a complex state, individuals use heuristic attributes, which substitute for the more complex state and are easier to assess.

13.4.2 A Response to the Small Improvements Argument from Bounded Rationality

It is not hard to see how heuristics can be employed in an epistemic solution to the Small Improvements Argument. "Life as a Lawyer" and "Life as a Philosopher" are heuristic attributes, which stand in for more complex states. When comparing *L* to *P*, one compares these two heuristic attributes and deems them the same in terms of their value. When comparing *L*+ to *P*, one still compares these *exact same* two heuristic attributes. Even if a fully specified fine-grained state of affairs of *L*+ is better than a fully specified fine-grained state of affairs *L*, we are not comparing either of those to *P*.[4]

Perhaps one reason why no one has understood the Small Improvements Argument in these terms is that Kahneman and Tversky and Gigerenzer are all explicit that heuristics are *fast and frugal*. But deliberation about careers is anything but fast and frugal, and so my framing the decision in terms of a heuristically mediated decision may seem misplaced. However, I think instead that we should view this use of a heuristic as what might be called a *slow heuristic*. The presumption in the Raz case (which I accept) is that even given all the time in the world, a normal person would still not be able to decide in advance about whether the career as a philosopher or a lawyer would be best. Nonetheless, we can easily recognize that we are already ignoring many small features in the possible future when comparing the heuristic state "life-as-a-lawyer" to the heuristic state "life-as-a-philosopher." And even when there is a distinct "life-as-a-lawyer+" to compare, we nevertheless still use the exact same heuristic state "life-as-a-lawyer" when comparing *L*+ to "life-as-a-philosopher." And that is the case even though when comparing the two lives-as-lawyers just to each other, we can easily make a comparative judgment to show that the *L*+ is slightly better than *L*. (To clarify, I am calling the use of "life-as-a-lawyer" and "life-as-a-philosopher" *slow* heuristics because they have involved a significant amount of thought; the deliberator does not generate and employ them quickly, as is typical for other heuristics such as an *availability* or *recognition* heuristic.)

This still leaves open the question of whether it is *irrational* to use a slow heuristic in these cases. What I'd like to suggest is a particular explanation of how the small improvement case works. In considering *L* and *L*+, what are the differences in the choice situations when comparing each to *P*? Why is it reasonable to not change one's heuristic when shifting from *L* to *L*+ in the choice scenario when one is comparing each to P, but it *is* rational to prefer *L*+ to *L*?

When one is comparing *L* to *P*, one is already ignoring many details of both situations; we must suppose that *some* ignoring of details is rational. The difference between *L* and *L*+ is smaller than features of *L* that the agent is already (by stipulation *rationally*) disregarding when considering *L*. It is plausible to think that what makes it rational to not shift one's heuristic when the opportunity for *L*+ becomes available is something like the following heuristic principle, which I will call the *Principle of Comparable Disregard (PCD)*:

> **PCD:** If, in generating a heuristic judgment, one is already rationally disregarding fine-grained consideration *C*, then it is rational to disregard other considerations of equal or lesser weight than *C*.

The PCD is a heuristic both in the familiar sense that it is a rule of thumb, but also in the more technical sense that it recommends substituting a simpler attribute (or set of attributes) for a more complex one for the sake of making deliberation more manageable.

I should say that the PCD is *not* universally valid. That's because enough iterations of it could allow a situation to pass a threshold whereby the new consideration *does* make a moral difference. But its point is not to be a universally valid principle. In *most* cases, the PCD is an appropriate heuristic principle to employ when considering options about which one lacks full knowledge, and one must make a decision on the basis of such incomplete knowledge. So, I want to emphasize the practical role that the PCD is playing. It is (1) a rule of thumb that people likely employ (even if non-explicitly) in the face of having too many considerations to take into account while making a decision in a context, and it is (2) reasonable to employ in virtually all contexts.

There is a long history within utilitarian theory that relates to this issue, though it has typically not been discussed in the vocabulary of *heuristics*. For the utilitarian, individuals ought to maximize the good, but it is not the case that individuals ought to spend their time thinking about how to maximize the good. Utilitarians have long held that utilitarian theory is a theory of right action rather than a decision procedure (see Bales 1971). R.M. Hare famously argues (1981, Part I) for *two-level utilitarianism*, where it is appropriate for individuals to make ordinary decisions at an intuitive level, only going to a critical level when necessary. As Jeremy Bentham writes, "It is not to be expected that this process [of utilitarian calculus] should be strictly pursued previously to every moral judgment .... It may, however, be always kept in view: and as near as the process actually pursued on these occasions approaches to it, so near will such process approach to the character of an exact one" (1789, Chap. IV, Sec. VI; see also Sinnott-Armstrong 2021, §4). On perhaps the best recent version of a two-level view, Fred Feldman (2012) gives an account of a utilitarian decision procedure in cases in which one does not know how to maximize utility, and the view here should be seen as fitting within Feldman's general framework. (Additionally, Yetter Chappell 2019, 105, recommends a consequentialist use of heuristics, and Armendt 2019 does so in defense of causal decision theory.) I will have more to say about this below, but the point for the moment is that principles such as the PCD, which can be used in conjunction with slow heuristics, can provide a bridge between one's inability to perform a full-fledged expected utility calculation under conditions of boundedness while still "keeping in view" the general idea of maximizing expected utility.

13.4.3 An Application of a Two-Level Account

An analogy with cigarette smoking may be helpful in showing the usefulness of the kind of two-level view I have in mind. What is the impact of smoking a *single* cigarette? The causal relationship between cigarette smoking and cancer is small, diffuse, and probabilistic and still not fully understood. Smoking can cause harms and lower life expectancy through multiple channels, and some people who smoke live long and quite healthy (and cancer-free) lives. The impact of one individual cigarette is tiny (at least for someone who is already a smoker). Is there nothing negative (from the perspective of one's own long-term self-interest) in smoking any single cigarette?

Kingston and Sinnott-Armstrong (2018) argue that the causal relationship between individual GHG emissions is tiny, diffuse, probabilistic, and in general best seen at levels of explanation higher than the individual level. At the same time, the formation of cancer is arguably emergent (see Plutynski 2018, Ch. 1 for discussion of the complexities in causal attribution of cancer) in a way analogous to how Kingston and Sinnott-Armstrong claim that climate change is emergent; and if so, then on Kingston and Sinnott-Armstrong's reasoning, smoking individual cigarettes does not cause cancer, and then there would be no reason not to smoke any given individual cigarette. However, this reasoning can be *iterated perpetually*, every time one considers having a cigarette. Of course, smoking is addictive in a way that driving is not (though driving arguably might positively correlate with driving on later days, as I shall note below), but the point remains that insofar as there is a choice involved in smoking individual cigarettes, there would be no reason to not smoke, *if* it were true that tiny or causally diffuse risks can always be discounted.

On the other hand, while a Kingston/Sinnott-Armstrong style argument might entail that one can reasonably discount individual cigarette smoking risks and thus show too much, a two-level view does not. It seems reasonable to say, in accord with an expected value approach, that every cigarette increases one's chances of health complications by some small amount; in a poignantly titled research letter ("Time for a Smoke? …"), Shaw et al. (2000) argue that every cigarette reduces life expectancy by 11 minutes. Furthermore, for those who smoke, it does not seem that *other* risks are already being ignored in the smoking of an individual cigarette that exceeds the health risks of smoking. So, a two-level view, supplemented with PCD, provides no reason to think that the risks of smoking should be ignored in deliberations on whether to smoke any particular cigarette. Maybe this is intuitively the right outcome. Or maybe not – perhaps some *other* heuristic considerations can be used to show that it is reasonable to ignore the risk of smoking some individual cigarettes on some special occasions, but whatever *those* heuristic grounds are, given the fact that the risks still aggregate on the negative side of the expected value ledger, the risk of a lifetime of cigarette smoking is not something that it is reasonable to ignore. (I will say more about lifetime decisions in §6.1.)

13.5 How to Think about Climate Risks from Individual GHG Emissions

How does the PCD apply in the case of individual actions that emit GHGs? For one thing, in the particular case of driving, we already know that there are risks involved. According to the United Nations, there are 1.3 *million* deaths and 50 million injuries annually from traffic accidents (2021).[5] On the face of it, in choosing to go for a drive, one is already ignoring risk, or at least not letting risk overwhelm other factors. I will say more about this shortly, but one might suppose at the outset that if it is reasonable to ignore the risk of direct road death, then it is also reasonable to ignore the risk of climate-related risk from driving. Perhaps people don't quite *ignore* risks from driving, as people do take some precautions; but still, people choose to drive in the face of risks, and for practical purposes typically don't bother to consider driving risks as playing any role in particular decisions whether or not to drive somewhere. If that's the case, and if, intuitively, climate risks are less significant than driving risks, then PCD would reasonably permit individuals to ignore them as well.

What does the most recent research say about the expected climate impacts of individual GHG-emitting life activities? Broome (2019) argues persuasively that the expected harm from individual GHG emissions is positive, and Broome (2021) uses the data compiled in Carleton et al. (2019) to arrive at average lifelong harm for a person in a developed country. On a low-estimate model in Carleton et al. (2019) of overall expected harm, Broome (2021) argues that individuals are responsible for approximately six *months* of harm to others; on a high model in Carleton et al. (2019), individuals are responsible for six-to-seven *years* of harm.[6]

Although I endorse most of Broome (2019, 2021), Broome's calculations are not in fact a full employment of expected value theory. Broome discusses the impacts of individual GHG emissions, but Broome's calculation does not determine the *marginal* effects of one's GHG-emitting activities. Just because one's action can be expected to contribute, say, .001% to a harm of *x* units, it doesn't follow that one is morally responsible for 1/100,000*x* units of harm. That's because what matters, according to expected value theory, is not one's personal contribution, but one's expected *marginal* contribution – the expected *difference* that one makes. For instance, if one chooses to not purchase a tank's worth of gas at a gas station and thus not emit particular carbon atoms into the atmosphere, someone *else* will most likely emit those very carbon atoms. So, the relevant questions are: how much of a difference to overall *net* global emissions is it expected to make when one increases or reduces one's individual emissions by a certain amount?[7] And once that is determined, how much of an expected difference in climate harm does that *difference* make?

This is the kind of issue that often arises in regard to inefficacy arguments in animal ethics. There is a difference between arguments regarding individual GHG emissions and arguments regarding consuming animal products. In the case of eating meat, the animal consumed is already dead, and so insofar as the wrongness of eating meat consists in causing harm, the wrongness must somehow be due to the ways in which one's meat purchase has incentivized the future harm to other animals via market mechanisms. On the other hand, the direct causal effect of one's GHG emissions has the potential, at least, to marginally increase global warming. For this reason, Broome believes that it is wrong to emit GHGs – there is a risk that one's very emissions will play a causal role. But it should be noted Broome is explicitly *not* using utilitarian reasoning when discussing individual obligations (see 2012, Ch. 4, 2021, 290). Broome is concerned with the very particles that one oneself is responsible for emitting, regardless of whether others would have emitted them.

This is relevant, for example, because in the time of measures to restrict movement due to the COVID pandemic, many bemoaned that GHG emissions were not reduced by as much as one would have hoped (see Tollefsen 2021). For example, some airlines continue to fly airplanes even with few or no passengers – so-called ghost flights – simply to preserve airport slots,[8] which gives evidence that even if consumers choose to reduce their personal emissions, it will not have an equivalent effect on net emissions reduction (though it should be noted that the percentage of ghost flights is still said to be "minute" relative to overall flights; also see Jiang et al. 2021 for a detailed and not-pessimistic analysis of the relationship between COVID and emissions).

Economists study *demand* and *supply elasticities*; this means, respectively, the changes in consumption, or production, of a product when the price of it goes up or down. In general, fossil fuel supply is rather *in*elastic in the short-medium term; there is a long supply chain between extraction of coal or oil and consumption, and short-term changes in demand won't have a direct effect on short-term production. For instance, Green and Denniss (2018, §3.3) discuss the phenomenon of infrastructure "lock-in." Producers make large up-front investments in production capacity, like investing in coal mines or oil fields. Even if overall consumer demand goes down, sending prices below the average level where the overall long-term investment in the infrastructure is profitable, the producer may still continue to produce and sell the product at the low amount. If the producer is already locked into the up-front investment in the infrastructure, continuing to sell the product at that price may still be a current marginal gain.

Furthermore, how an increased supply translates into increased GHG emissions is complex. It may drive prices down, driving consumption up. However, demand, on the whole, is also fairly inelastic relative to price in the short-medium term[9]: this is because people have regular schedules in which they commute to work, make travel plans, etc., and in general do not change their fossil fuel consumption significantly in direct response to price changes. The relative demand inelasticity of fossil fuels means that a scenario implied by Johnson (2003), where some people start forgoing a limited resource, leading to lower prices, which in turn incentivizes others to increase consumption of the resource – would not be a significant concern for fossil fuels. Over the long term, however, both supply and demand are indeed somewhat sensitive to changes in price (see Güntner 2014; Krichene 2002), which means that long-term reductions in consumption can indeed be expected to lead to long-term reductions in supply.

If one individual refrains from buying a tank of gasoline and emitting the GHGs in it, those GHG molecules will instead be emitted by someone else, but that next person would then not emit what would otherwise have been their own tank's worth of GHGs, which would then be emitted by someone else, and so on. However, setting aside larger macroeconomic factors, the first person's restraint means that, at any given point in time, net GHG emissions will be lowered by that amount. Once we include macroeconomic factors, perhaps the amount of difference in net emissions will be less than the amount that the individual has refrained from emitting, but all in all, there is little reason to believe that over either the short or long run, the difference in overall emissions is insensitive to individual reductions in consumption.

There are a couple of upshots of the considerations from the economists and from Broome in this section. First, there is a great deal of uncertainty, still, about the *marginal* effects of individual actions, even lifetime individual actions, and that, despite excellent efforts from Broome, our best estimates may still not represent the true range of possibility of climate-related effects, and that this is a reason for further investigation on the question. Economists are interested in market changes due to changes in consumer activity, and ethicists are interested in expected values of decisions, and while ethicists can look to economists for clues, because the questions are not the same, the empirical work from economists does not always directly address the questions to which ethicists wish for answers. That being said, it seems that because of the inelasticity of demand in the short-medium term, and increased elasticity of supply in the long term, it is not unreasonable to think that the marginal effects of GHG-emitting activities do not depart *dramatically* from the average effects as calculated by Broome (2021).

Finally, we can return to the question with which I began this section: how does the PCD apply to climate risks from driving? According to the United States National Highway Traffic Safety Administration (see undated "Summary Table"), there were approximately 1.11 traffic fatalities per 100 million vehicle miles traveled in the United States in 2019. According to the United States Environmental Protection Agency, "The average passenger vehicle emits about 404 grams of CO2 per mile" (2021), or in other words, 40,400 metric tonnes of CO_2 per 100 million miles. Using Broome's (2021) lower harm estimate, this amounts to approximately 17 life-years per 100 million miles, and on the higher harm estimate, it amounts to approximately 220 life-years.[10] This suggests that the risk of causing traffic fatalities is just above the expected harm from a car's GHG emissions in the low-harm estimate, and quite a bit below it in the high-harm estimate.

I confess that these are still all rather rough estimates, and I hope that in time, better estimates will be developed. Furthermore, given the above considerations regarding the difference between the average harm and the marginal harm from GHG emissions, I am unsure exactly how to modify these values in light of market inelasticities, and I encourage the reader to come to their own conclusions in light of the details here. My own sense is that while there is still significant uncertainty about the level of climate-related risk of individual actions, it would nevertheless not be reasonable to argue on the basis of the considerations above that it is *clearly* the case that climate risks from driving are less significant than direct traffic fatalities, and to claim that these risks can be fully disregarded. On the one hand, the point of using a heuristic is to avoid having to do these kinds of calculations in the first place. On the other hand, the information I have provided here can be employed by those who may wish to generate a *slow* heuristic using a general "approximate climate risks per action." But this kind of heuristic does not seem to give reason, on its own or using the PCD, to *ignore* climate risks, even given all the uncertainties involved.

I'd like to end this section by clarifying several open issues. First, I have focused on PCD as a heuristic principle; perhaps there are *other* heuristic principles that can be reasonably employed that make it reasonable to ignore climate-related risks of driving. The PCD cannot be the only applicable heuristic principle, since it itself makes reference to fine-grained considerations *already* being properly ignored. To that extent, my argument is limited. But I should note that even if there are such other heuristic principles that permit the ignoring of climate risk, the upshot would be different than that of typical inefficacy arguments; rather than agreeing that climate risks make no difference in ordinary actions, the most one could say would be that (within a two-level framework) it is sometimes reasonable to ignore them. (The same might be said for ignoring risks from individual cigarettes.) Second, I should also note, in accordance with Ori (2020), that we should perhaps also not be so quick to ignore traffic-related risks from driving. If so, then the case for not ignoring climate risks becomes even stronger. Third, I haven't discussed *other* ordinary

GHG-emitting activities. But my hope is that my discussion here of the economics of climate risks and of heuristics (and PCD in particular) can be modeled to apply to other cases as well.

13.6 Beyond Individual Actions

13.6.1 A Note on Single-Action vs. Lifespan Decision-Making and Planning

There are certainly many actions whose GHG-related expected disvalue is so small that the PCD indeed likely applies. For instance, when one boils water for tea, one emits GHGs, but one also runs the risk of burning oneself on the pot or the water itself. What does the view in this paper say about such actions? Furthermore, is there a meta-level at which we must engage in a determination to see if PCD applies? The whole point of PCD is to avoid having to over-calculate in particular circumstances.

To be clear, I accept that there is still some small expected disvalue for such actions to be placed on the negative side of the ledger. However, it is still the case that for many such actions, it is reasonable to ignore the GHG-related effects. (As Sinnott-Armstrong 2005 points out, even breathing emits CO_2; a view that held that one must constantly keep climate change in mind in each breath is absurd.)

While much of the literature regarding individual climate ethics has viewed GHG-emitting choices as independent events, as Michael Bratman has long pointed out (cf. 1987), our lives are not a series of disconnected choices but rather are structured by plans. Dale Jamieson (2007) also notes that, regarding climate change, utilitarians ought to be virtue ethicists, so as to instill durable pro-environmental character traits. And Marion Hourdequin (2010) and Trevor Hedberg (2018) argue that, on grounds of integrity, if one cares about the environment, one should try to refrain from GHG-emitting activities even if the effects are limited. This also relates to the analogy with smoking: it can be argued that the proper way to view smoking decisions is not as decisions to smoke individual cigarettes, but to purchase packs or cartons, or to quit this month or not to quit. And it is these decisions that have more impact than individual cigarette choice risk. Perhaps driving is similar, insofar as one chooses whether to have a car, or where to live relative to one's job (although it should be noted these choices are often constrained in one way or another by financial limitations), and these decisions have larger impacts on overall GHG emissions than decisions about whether or not to go on individual Sunday drives. (One may also decide on general hobbies for one's days off – one can choose to be the kind of person who travels somewhere distant most weekends or who does activities close to home.) My point here is just that from the perspective of individual decision-making with regard to climate risk, there are a number of relevant levels of analysis, and the level of analysis of the individual Sunday drive, while not inappropriate, is perhaps not the most important one when considering the individual risk of our decision-making. In this way, there is some truth to taking a broader perspective like in Kingston and Sinnott-Armstrong (2018) and Gesang (2021) – but that doesn't show that *single* individual actions make no difference.

Taking some time at various stages of one's life is consistent with Bentham's suggestion (and Hare's two-level view) to not always focus on abiding by expected utility theory but still keep in mind the overall goals of maximizing utility. The best level of assessment for individuals with regard to climate change is at the level of the individual's more general life-plan. Individuals, especially those in wealthy nations who are more than capable of doing so, should structure their lifestyle so as to reduce or limit activities, even small-scale ones, that emit GHGs, though of course in our era it is impossible to eliminate them entirely.

13.6.2 On the Relation between Individual and Group Action

One might wonder whether it is frivolous to once again discuss individual action in the contexts of climate change – one often hears claims that we must overthrow the system and not dwell on little things. I would like to make several points in response. First, as I myself have noted (2011b, 365), creating political change faces some of the same inefficacy concerns as reducing climate change. Second, it should be emphasized that telling people in a public forum that individual decisions make a difference (or do not make a difference) is *not* an individual action – it is a collective action. Third, one sometimes hears claims that individual changes will not *stop* climate change. As the headline of an article in Time Magazine by climate scientist Michael Mann puts it (2019), "Lifestyle Changes Aren't Enough to Save the Planet." But the truth is that it is now impossible to *stop* climate change. And there has never been a question of "saving the planet." What it is not too late to do is incrementally lessen the negative impacts of climate change. The rhetoric of "stopping climate change" or "saving the planet" is inappropriate and perhaps leads people to reject the expected value approach and its incrementalism. The problem here is with the rhetoric, and not with expected value theory.

Sometimes, ignoring decimal places in a numerical claim is conversationally appropriate, according to Gricean norms of conversation (Grice 1989). Telling someone that it is 9:01:17 is conversationally *worse* in many contexts than just saying that it is nine o'clock. Likewise, telling someone about to go for a short drive that their GHG emissions make no difference may, in some conversational contexts, be more conversationally cooperative than saying that it makes a tiny difference. But if the considerations in this paper are correct, then a *general* practice of telling people that short drives make no difference, across conversational contexts, is not appropriate. Perhaps the claim that individual emissions make no difference can be used to *flout* Gricean quality norms – of course individual GHG emissions have *some* expected disvalue! – so the implicature goes that the disvalue is so small that we should focus on other things, like changing laws or overthrowing the system.

Those are admirable socio-political goals. And philosophers have long discussed *collective* responsibilities (see, e.g., the essays in Bazargan-Forward and Tollefsen, 2020) – a topic I have not broached here. But we can *both* be mindful of our emissions – sometimes, if the view in this paper is correct, in the back of our minds – and also mindful of the best ways to engage in political and other collective activities to best limit climate change. And we should act upon those intentions, both in our personal and – insofar as the personal is not *already* political – in our public and political lives.

13.7 Conclusions

To sum up, I have argued that although it is reasonable to ignore some risks, one ought, at various points in one's life, to consider the climate impacts of one's lifestyle and attempt to formulate life-plans in ways that take into account the expected harms of one's GHG emissions. I have argued that *some* small risks may reasonably be ignored because human psychology is bounded not just by the amount of time we have to make decisions but also by the near-impossibility in many cases of acquiring conclusive evidence about the details of future states of affairs that may ensue if one chooses an option. I have further argued (in accord with a long line of two-level ethical theorists) that these claims are not in violation of the spirit of expected value theory, which accepts that there are costs to calculating risks in particular cases and thus does not require individuals to do so constantly. Nevertheless, even the reasonable use of heuristics does not seem to permit one to ignore the climate-related risks of individual actions (although the climate risks of certain of one's actions may still be outweighed by their positive expected effects) and does not seem to permit the general condoning of ignoring individual actions' climate risks.

My aim in discussing heuristic principles is to give voice to something correct that may underlie some individual inefficacy concerns – that sometimes, risks may be too insignificant to merit consideration – without undermining the view (such as in Broome 2019; Hiller 2011a, 2011b) that there is still some amount of expected disutility of ordinary actions due to climate risks (and that this disutility is likely not miniscule). Why does this matter? One might wonder who bears *responsibility* for making changes to help reduce climate risk. Perhaps ordinary individuals, in deciding upon *some* minor actions, may be reasonable in ignoring climate-related risks, but the fact that there is always going to be some amount on the negative side of the ledger due to climate risk means that we as individuals do indeed bear some moral responsibility for mitigating the climate risks we impose upon others (in addition to there being collective responsibility for climate risks as well). And as I have tried to argue, here and elsewhere, these risks are not so small as to be morally insignificant. Inefficacy arguments leave no room for this assessment.

So much of the nitty-gritty of life involves doing little things that make little (expected) difference. One might walk one more block for exercise; one might slightly lower the pitch of one's voice in conversation with a friend in need of calm; one might add a shake of garam masala to the pot of dal one is cooking; one might add a touch more vibrato on a fifth note in a guitar riff one is playing; one might take a multivitamin; one might carry a sign in a protest line; one might make a kind facial expression to a fellow passenger on the bus.

The general perspective from which the denial of the expected utility approach arises – that small actions make no difference – may fully nullify the significance of these actions and thus potentially leave one no reason for doing them. Perhaps advocates of inefficacy arguments can, for each of these domains, provide independent reasons for claiming that these small actions are appropriate or not, despite their making no difference (or no noticeable one). But the vast heterogeneity of cases in which small-scale actions occur, and to which we intuitively attach value judgments, suggests that a general solution is probably the best one. And the general solution given by expected value theory – that small-scale actions do indeed make small *expected* differences, which then can add up – for (expected) better or (expected) worse – is the most plausible and theoretically elegant way of accounting for the value of these actions, even if expected value theory is nevertheless not always action-*guiding* because of the boundedness of human capacities.[11]

Notes

[1] This debate echoes one in animal ethics, where some philosophers argue that we do not have an obligation to reduce our meat consumption despite the fact that animals are deserving of moral consideration; see, e.g., Fischer (2019, Ch. 4) and Nath (2021).

[2] It should be noted that this notion of discounting differs from *temporal discounting*, a common notion within climate ethics.

[3] Broome (2019, 124) has some discussion of the Small Improvements Argument, but Broome's use of the SIA is quite different from the one in this paper.

[4] This analysis of the Small Improvements Argument is similar to that in Anderson (2015).

[5] See Ori 2020 for more on road ethics.

[6] In my own work (Hiller 2011a, 2011b, 2014), I use an estimate from John Nolt (2011) according to which one individual is responsible for one lifetime's worth of harm; Broome's (2021) estimates now seem plausible to me, though I have some concerns with Broome's analysis, both for reasons I discuss below, and also due to issues of harm to non-humans. Additionally, in Hiller (2011a, 2011b), I begin (349) with the question of how much harm individual GHG-emitting acts *cause*. Gunnemyr (2019) criticizes the claim that individual GHG-emitting acts *cause* harms, and I agree with much of Gunnemyr's critique. But I wish to emphasize here that for expected value theory, it is not *causing* harm, but *expected* harm, that matters. (Also, in Hiller [2011a, 2011b], I misleadingly use the phrase "causes an expected harm" [355]; this phrase is infelicitous. One does not *cause* an expected harm with an action. Rather, actions *have* expected harms.)

[7] For a discussion of this point, see Hale (2011). As I argue in Hiller (2011b), the *timing* of emissions matters, so even if an earlier GHG emission is replaced by an equivalent later emission, the two will not be equal in their climate-related effects.

[8] See, for example, Chris Stokel-Walker, "Thousands of Planes Are Flying Empty and No One Can Stop Them," *Wired*, February 2, 2022, https://www.wired.com/story/airplanes-empty-slots-covid/.

[9] See Michael Morris, "Gasoline prices tend to have little effect on demand for car travel," *U.S. Energy Information Administration*, December 17, 2014, https://www.eia.gov/todayinenergy/detail.php?id=19191.

[10] The lower estimate in Broome (2021) is half a year of harm per 1200 tonnes; solving for $\frac{.5}{1200} = \frac{x}{40,400}$ yields x ≈ 17. The higher estimate is six-to-seven years for the same amount of emissions; solving for $\frac{6.5}{1200} = \frac{x}{40,400}$ yields x ≈ 220.

[11] Many thanks to Adriana Placani and Stearns Broadhead for extremely helpful comments on an earlier version of this chapter.

References

Adler, Matthew D. 2007. "Why De Minimis?" *University of Penn, Institute for Law & Econ Research Paper*, No. 07–12: 7–26.

Anderson, Jack. 2015. "Resolving the Small Improvement Argument: A Defense of the Axiom of Completeness." *Erasmus Journal for Philosophy and Economics* 8 (1): 24–41.

Andreou, Chrisoula. 2015. "Parity, Comparability, and Choice." *Journal of Philosophy* 112 (1): 5–22.

Armendt, Brad. 2019. "Causal Decision Theory and Decision Instability." *Journal of Philosophy* 116 (5): 263–77.

Bales, R. Eugene. 1971. "Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedures?" *American Philosophical Quarterly* 8: 257–65.

Balfour, Dylan. 2021. "Pascal's Mugger Strikes Again." *Utilitas* 33 (1): 118–24.

Barrington, Mitchell. 2021. *Ignoring the Improbable*. Unpublished Manuscript. http://www.mitchellbarrington.com/wp-content/uploads/2022/01/Ignoring-the-Improbable.pdf

Bazargan-Forward, Saba, and Deborah Tollefsen (Eds.). 2020. *The Routledge Handbook of Collective Responsibility*. New York: Routledge.

Bentham, Jeremy. 1789/2007. *An Introduction to the Principles of Morals and Legislation*. New York: Dover Press.

Bostrom, Nick. 2009. "Pascal's Mugging." *Analysis* 69 (3):443–45.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Broome, John. 2012. *Climate Matters: Ethics in a Warming World*. New York: W.W. Norton.

Broome, John. 2019. "Against Denialism." *The Monist* 102: 110–29.

Broome, John. 2021. How Much Harm Does Each of Us Do? In *Philosophy and Climate Change*, edited by Mark Budolfson, Tristram McPherson and David Plunkett, 281–92. New York: Oxford University Press.

Carleton, Tamma, Michael Delgado, Michael Greenstone, Trevor Houser, Solomon Hsiang, Andrew Hultgren, Amir Jina, Robert Kopp, Kelly McCusker, Ishan Nath, James Rising, Ashwin Rode, Samuel Seo, Justin Simcock, Arvid Viaene, Jiacan Yuan, and Aice Zhang. 2019. "Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits." *BFI Working Paper*. Chicago: Becker Friedman Institute for Economics at the University of Chicago.

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112 (4): 659–88.

Chappell, Richard Yetter. 2019. "Fittingness Objections to Consequentialism." In *Consequentialism: New Directions, New Problems?*, edited by Christian Seidel, 90–111. Oxford: Oxford University Press.

Cripps, Elizabeth. 2013. *Climate Change and the Moral Agent: Individual Duties in an Interdependent World*. Oxford: Oxford University Press.

De Sousa, Ronald B. 1974. "The Good and the True." *Mind* 83 (332): 534–51.

de Staël, Madame (Anne-Louise Germaine). 1845. *De L'Allemagne*. Paris: Didot Frères.

Ebert, Philip A., Martin Smith, and Ian Durbach. 2020. "Varieties of Risk." *Philosophy and Phenomenological Research* 101 (2): 432–55.

Fischer, Bob. 2019. *The Ethics of Eating Animals: Usually Bad, Sometimes Wrong, Often Permissible*. New York: Routledge.

Gesang, Bernward. 2017. "Climate Change—Do I Make a Difference?" *Environmental Ethics* 39 (1): 3–19.

Gesang, Bernward. 2021. "Utilitarianism and Heuristics." *Journal of Value Inquiry* 55 (4): 705–23.

Gigerenzer, Gerd. 2000. *Adaptive Thinking: Rationality in the Real World*. Oxford: Oxford University Press.

Gigerenzer, Gerd. 2021. What Is Bounded Rationality. In *Routledge Handbook of Bounded Rationality*, edited by Riccardo Viale, 55–69. New York: Routledge.

Green, Fergus, and Richard Denniss. 2018. "Cutting with Both Arms of the Scissors: The Economic and Political Case for Restrictive Supply-Side Climate Policies." *Climatic Change* 150: 73–87.

Grice, H. Paul. 1989. Logic and Conversation. In *Studies in the Way of Words*. Cambridge: Harvard University Press.

Gunnemyr, Mattias. 2019. “Causing Global Warming.” *Ethical Theory and Moral Practice* 22 (2): 399–424.

Güntner, Jochen H. 2014. “How Do Oil Producers Respond to Oil Demand Shocks?” *Energy Economics* 44: 1–13.

Hale, Benjamin. 2011. “Nonrenewable Resources and the Inevitability of Outcomes.” *The Monist* 94 (3): 369–90.

Hare, Richard M. 1981. *Moral Thinking*. Oxford: Clarendon Press.

Hedberg, Trevor. 2018. “Climate Change, Moral Integrity, and Obligations to Reduce Individual Greenhouse Gas Emissions.” *Ethics, Policy and Environment* 21 (1): 64–80.

Hiller, Avram. 2011a. “Morally Significant Effects of Ordinary Individual Actions.” *Ethics, Policy and Environment* 14 (1): 19–21.

Hiller, Avram. 2011b. “Climate Change and Individual Responsibility.” *The Monist* 94 (3): 349–68.

Hiller, Avram. 2014. “A ‘Famine, Affluence, and Morality’ for Climate Change?” *Public Affairs Quarterly* 28 (1): 19–39.

Hourdequin, Marion. 2010. “Climate, Collective Action and Individual Ethical Obligations.” *Environmental Values* 19 (4): 443–64.

Jamieson, Dale. 2007. “When Utilitarians Should Be Virtue Theorists.” *Utilitas* 19 (2): 160–83.

Jiang, Peng, Yee Van Fan, and Jiří Jaromír Klemeš. 2021. “Impacts of COVID-19 on Energy Demand and Consumption: Challenges, Lessons and Emerging Opportunities.” *Applied Energy* 285: 116441.

Johnson, Baylor. 2003. “Ethical Obligations in a Tragedy of the Commons.” *Environmental Values* 12 (3): 271–87.

Kahneman, Daniel. 2003. “Maps of Bounded Rationality: Psychology for Behavioral Economics.” *American Economic Review* 93 (5): 1449–75.

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. New York: Macmillan.

Kahneman, Daniel, and Shane Frederick. 2002. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In *Heuristics and Biases: The Psychology of Intuitive Judgment*, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 49–81. Cambridge: Cambridge University Press.

Kingston, Ewan, and Walter Sinnott-Armstrong. 2018. “What’s Wrong with Joyguzzling?” *Ethical Theory and Moral Practice* 21 (1): 169–86.

Knight, Frank. 1921. *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin.

Kosonen, Petra. 2021. “Discounting Small Probabilities Solves the Intrapersonal Addition Paradox.” *Ethics* 132: 204–17.

Krichene, Noureddine. 2002. “World Crude Oil and Natural Gas: A Demand and Supply Model.” *Energy Economics* 24 (6): 557–76.

Lundgren, Björn, and Hlynur Orri Stefánsson. 2020. “Against the de minimis principle.” *Risk Analysis* 40 (5): 908–14.

Mann, Michael E. 2019. “Lifestyle Changes Aren’t Enough to Save the Planet. Here’s What Could.” *Time Magazine*. https://time.com/5669071/lifestyle-changes-climate-change/ (accessed 30 March 2022).

Monton, Bradley. 2019. "How to Avoid Maximizing Expected Utility." *Philosophers' Imprint* 19: 1–25.

Morgan-Knapp, Christopher, and Charles Goodman. 2015. "Consequentialism, Climate Harm and Individual Obligations." *Ethical Theory and Moral Practice* 18 (1): 177–90.

Nath, Rekha. 2021. "Individual Responsibility, Large-Scale Harms, and Radical Uncertainty." *The Journal of Ethics* 25 (3): 267–91.

Nefsky, Julia. 2012. "Consequentialism and the Problem of Collective Harm: A Reply to Kagan." *Philosophy and Public Affairs* 39: 364–95.

Nolt, John. 2011. "How Harmful Are the Average American's Greenhouse Gas Emissions?" *Ethics, Policy & Environment* 14: 3–10.

Nye, Howard. 2021. "Why Should We Try to be Sustainable? Expected Consequences and the Ethics of Making an Indeterminate Difference." In *Right Research: Modelling Sustainable Research Practices in the Anthropocene*, edited by Chelsea Miya, Oliver Rossier and Geoffrey Rockwell, 3–35. Cambridge: Open Book Publishers.

Ori, Meshi. 2020. "Why Not Road Ethics?" *Theoria* 86 (3): 389–412.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Peterson, Martin. 2002. "What Is a De Minimis Risk?" *Risk Management 4* (2): 47–55.

Plutynski, Anya. 2018. *Explaining Cancer: Finding Order in Disorder*. New York: Oxford University Press.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Oxford University Press.

Regan, Donald H. 1989. "Authority and value: Reflections on Raz' morality and freedom." *Southern California Law Review* 62: 995–1095.

Sandberg, Joakim. 2011. "My Emissions Make No Difference': Climate Change and the Argument from Inconsequentialism." *Environmental Ethics* 33 (3): 229–48.

Shaw, Mary, Richard Mitchell, and Danny Dorling. 2000. "Time for a Smoke? One Cigarette Reduces Your Life by 11 Minutes." *BMJ (Clinical Research Ed.) 320* (7226): 53.

Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* 69: 99–118.

Sinnott-Armstrong, Walter. 2005. "It's Not My Fault: Global Warming and Individual Moral Obligations." In *Perspectives on Climate Change*, edited by Walter Sinnott-Armstrong and Richard Howarth, 221–53. Bingley: JAI Press.

Sinnott-Armstrong, Walter. 2021. Consequentialism. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2021/entries/consequentialism/; accessed April 22, 2022.

Smith, Nicholas J. J. 2014. "Is Evaluative Compositionality a Requirement of Rationality?" *Mind* 123: 457–502.

Sunstein, Cass. 2002. *Risk and Reason: Safety, Law, and the Environment*. Cambridge: Cambridge University Press.

Tollefsen, Jeff. 2021. "COVID Curbed Carbon Emissions in 2020—But Not by Much." *Nature 589* (7842): 343–3.

United Nations. "With 1.3 Million Annual Road Deaths, UN Wants to Halve Number by 2030." *UN News*. December 3, 2021. https://news.un.org/en/story/2021/12/1107152; accessed April 22, 2022.

United States Environmental Protection Agency. "*Greenhouse Gas Emissions from a Typical Passenger Vehicle*." Last modified July 21, 2021. https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle; accessed April 22, 2022.

United States National Highway Traffic Safety Administration (Undated) "Summary Table", https://www-fars.nhtsa.dot.gov/Main/index.aspx; accessed April 22, 2022.

Index

Note: **Bold** pages refer tables and with “n” notes in the text.