FILOSOFISKA NOTISER

Årgång 5, Nr 1, Maj 2018

ISSN: 2002-0198

Hemsida: www.filosofiskanotiser.com

Thought Experiment as an Interdisciplinary Pedagogy

Sharon Kaye

Abstract
The thesis of this paper is that thought experiments provide an especially powerful way to frame a class discussion. They work for students for the same reason that they have worked for great geniuses (such as Einstein) through the ages—namely, because they are interdisciplinary. Competing rationalist and empiricist accounts of how thought experiments work suggest that they will engage both rationally- and empirically-minded students. Examples of student responses to thought experiments confirm that they bring out interestingly diverse ways of thinking. Concern that inter-disciplinary pedagogy makes genuine communication impossible has led some theorists to insist on a methodological pluralism that refuses to privilege any one approach. I argue however, that interdisciplinary instructors must ultimately ask students to incorporate their diverse perspectives into the discourse of the instructor's discipline in order to ensure that their work is judged in accordance with a time-tested criterion of excellence.

Introduction
I began one of my classes this semester with the following thought experiment:

> You came to this room expecting a lecture, but this is a sting. My name is Agent Dana Scully, I am with the FBI, and you are under arrest. Please hand over your wallets, your cell phones, and your belts.… What? You say you are innocent? Well, let me enumerate just a few of the laws you have broken lately: first, traffic violation, which you commit just about every time you drive; second, copyright violation—pirated movies, and music—did you know that Time Warner owns the rights to the Happy Birthday song?; third, substance abuse—need I say more?; fourth, tax fraud, … the list goes on and on…. Still not willing to turn yourself in? Let us take a vote. How many of you feel you are innocent? Why?

Filosofiska Notiser, Årgång 5, Nr. 1, Maj 2018, 3–17

The question this thought experiment produces, "Do you really think you have the right to live in a country whose laws you regularly disobey?" is the very question Socrates posed before drinking the hemlock with which his fellow citizens sentenced him to death (Plato, 2002a, 45-57). It launches an exploration of justice, a concept central to many courses in higher education, including philosophy, sociology, religious studies, English literature, to name a few.

Thought experiments are an especially powerful way to frame a class discussion. They work for students for the same reason that they have worked for great geniuses (such as Einstein) through the ages—namely, because they are interdisciplinary.

What is a thought experiment?

A thought experiment is an imaginary scenario explored for the purpose of acquiring knowledge. Since the beginning of Western civilization, thought experiments have been used with great success in almost every field, from science, to ethics, to history (Rescher, pp. 61–72, 2005).

The ancient Roman philosopher Lucretius provides a classic example: Imagine throwing a spear at the edge of the universe. Either it will keep on going or it will hit a boundary. If it keeps on going, then you are not at the edge after all. But if it hits a boundary, then you are not at the edge either because a boundary is a divider with something on the other side. Lucretius believed this thought experiment proved that there is no edge—the universe must be infinitely large (Bailey, 1950, pp. 58–59).

There have been many competing accounts of the nature of thought experiment throughout history (Weber, 2003, pp. 28–38; Perler, 2008, pp. 143–153. Brooks, 1994, pp. 71–83). More recently, Letitia Meynell argues that thought experiments are ultimately props for imagining fictional worlds (Meynell, 2014, pp. 4149–4168.). She identifies six distinctive features that are common among them. These features help us understand what a thought experiment is.

The first is imagery—whether in the form of a diagram or just a mental picture. Imagery serves to stimulate the imagination and to ensure that all parties to the conversation are focused on the same idea. Lucretius does not abstractly posit "a projectile," he posits a spear—a very common object in his day. In my classes, I use my best acting skills—to the point where students complain, "Stop! You're scaring me!" The more concretely one can encapsulate the questions at hand, the more successful the thought experiment.

The second is experiential language. Lucretius does not ask us to imagine someone else throwing a spear. He puts us right there in outer space, poised to find out firsthand what happens to the spear. By casting us as the agent, not just an observer, Lucretius engages the senses as well as the intellect. We feel the spear; we see it fly. Likewise, I want my students to feel the cold metal of the handcuffs on their wrists. Through make-believe, we are fully engaged, the better to illicit authentic intuitions about the situation.

The third is an epistemological analysis, showing how the thought experiment justifies (or fails to justify) its conclusion. Though it is difficult to explain how imaginary scenarios produce knowledge, it is clear that we gain insight by understanding how they relate to our beliefs. Lucretius imagined a universe that must either have an edge or must go on infinitely because he assumed that space is Euclidean or flat. Today, space is no longer regarded as Euclidean, but rather curved in a complex way. Hence the thought experiment no longer regarded as providing insight into the way the world actually is. But this was only discovered by unearthing the beliefs underlying the scenario. Needless to say, there are competing accounts of how the beliefs involved in thought experimenting become knowledge (Clatterbuck, 2013, pp. 309–329).

Fourth is the irreducibly imaginative character of most thought experiments. While it might be possible to restate Lucretius's case in purely propositional form, the same is not true of the FBI thought experiment from my classroom. This scenario is not so much trying to prove that you are not innocent, but rather that your assumption of innocence is problematic. In a similar vein, Lucretius's thought experiment was, and still is, successful in so far as it proves that your assumption that the universe is finite is problematic. By serving up the problem rather than stating a conclusion, a thought experiment opens up conceptual space. This "laboratory of the mind" is characteristic of novels, plays, and other forms of fiction (Elgin, 1993, pp. 13–28; 2007, pp. 43-54; and 2014, 221–241).

Fifth, thought experiments tend to admit of different interpretations and to provoke opposition. For example, Aristotle objected to Lucretius's conclusion on the grounds that the world could not rotate uniformly if it were infinite in size. His geometrical proof of this claim is highly abstract and hence never became as famous as its rival. A more modern objection along the same lines might be: How can the universe be infinite if it is expanding? At any rate, thought experiments are designed to provoke thoughts—to raise more questions than they answer.

Finally, and most significantly, thought experiments are objective even though they are not real. When people discuss a thought experiment, it is

crucial that they agree to the "ground rules" of their make-believe world. If they imagine whatever they please and don't fully reveal to one another what they are imagining, then they make no progress on the question at hand. As though playing a video-game together, they must construct a virtual reality of fictional truths. Disagreements often arise from different rules, and insights are often gained by bringing to light hidden or unspoken rules. Rather than being private and subjective, thought experiments specify distinctive cognitive content. In this way they have "being in their non-being" (Meinong, 1907, 273–283).

How do thought experiments work?

Thought experiments are puzzling because we do not ordinarily think of imagination as a tool for knowledge acquisition. On the contrary, imagination, a fantasy about that which is not real, is commonly considered the opposite of fact, that which is real. How can a fantasy produce reliable information about reality?

Two competing answers to this question have emerged in the last ten years. We should consider them each in turn.

James R. Brown proposes that finely tuned imagination is actually a powerful form of mental perception. We all agree that physical perception—sight, hearing, smelling, touching, and tasting—is a way of gathering data about the physical world. Likewise, mental perception gathers data about truths that transcend the physical world (Brown, 2004).

Brown characterizes his view as "Platonic" with reference to the ancient Greek philosopher Plato. One of Plato's lifelong concerns was to identify a reliable source of truth. Rejecting physical perception as unreliable, he turned to mathematics as an ideal model. When we contemplate the equation a2+b2=c2 we "see" the truth with the mind's "eye." This seems to imply that there is another world, beyond the physical world, for us to discover. Plato called it the world of Forms (Plato, 2002b, pp. 91–3). He believed that human beings must have had access to this world before we were born since perceiving it feels like remembering.

For Brown, thought experimenting can provide a highly effective form of mental perception. When Lucretius imagined himself throwing a spear at the edge of the universe, he discovered something true that he could never perceive physically. This suggests that mental perception is not limited to math but can be extended to any area of inquiry (Brown, 2010, pp. 1–15). Brown's account is called "rationalist" because it holds that human beings can acquire knowledge through pure reason, without depending on physical perception.

John D. Norton presents the opposing empiricist account of how thought experiments work. In Norton's view, Plato's world of Forms does not exist. All knowledge comes either directly or indirectly from physical perception. Even mathematical equations such as a2+b2=c2 have an empirical source: they are abstracted from our observations of physical objects. Likewise, when Lucretius imagined throwing a spear at the edge of the universe, he extrapolated from his real life experience (Norton, 2004b, pp. 44–66).

Norton maintains that every thought experiment is really an argument in disguise (Norton, 2004a, pp. 1139–1151). Lucretius's thought experiment, for example, may be reconstructed as follows:

1. If the universe is finite, then it is surrounded by a final boundary.
2. But no boundary can be final because there always has to be something on the other side.
3. Therefore, the universe must be infinite.

We can picture Lucretius using his thought experiment to convince his opponents of his conclusion.

Norton contends, against Brown, that there is no reason to suppose Lucretius's thought experiment helped him to discover his conclusion. Surely the discovery came through reflection on ordinary empirical observation of various kinds of physical boundaries. The human mind is able to collect data from repeated experience and then construct abstract representations of things it is unable to experience. For example, we construct the idea of a perfect triangle by abstracting imperfections from the various physical triangles we encounter every day. Like Plato's illustrious student Aristotle, Norton insists that there is no justification for supposing that human beings can remember a transcendent world in which such truths exist. The mind is born a blank slate (Aristotle, 1986, 3.4.430a1).

Brown and Norton occupy opposite ends of the spectrum in explaining how thought experiments work. For Brown, thought experiments focus the mental perception that enables humans to discover transcendent truths. For Norton, they provide convincing illustrations for arguments rooted in empirical observation. I maintain, however, that these rival theories are not mutually exclusive from a pedagogical point of view.

The first clue to the underlying value of thought experiment comes from noticing that they very often don't produce true conclusions at all. Lucretius's spear, for example, doesn't actually prove that the universe is infinite at all.

> This thought experiment fails because the thought experimenter has overlooked the fact that it is actually possible for a surface to both be finite and have no edge; the surface of a sphere is an example. The thought experimenter mistakenly saw a contradiction when there is none (Cooper 343).

Newton's thought experiments overturned those of his ancient predecessors; Einstein's overturned Newton's, and recent thought experiments in quantum mechanics overturned Einstein's. Science continues to progress as do all other fields of inquiry. We have not settled upon the final answer.

And yet, each moment of overturning is flash of brilliant insight. If we do not learn the final answers in those moments, exactly what do we learn?

In those moments we learn about how we think. We see the power of a particular line of logic. We see the workings of the human mind at its best. While this thinking about thinking plays out with searing intensity at the professional level, it can be profitably modeled among amateurs in the classroom. As Elke Brendel writes,

> the long and sometimes fruitless debate in epistemology between internalist and externalist approaches to knowledge could indicate that there is not just one single concept of knowledge but at least two different concepts, each of which reflects different features of knowledge…. With the help of thought experiments these divergent, but legitimate concepts of knowledge can be clarified (2004, p. 104).

In a similar vein, Jeremy Goodenough documents how a single thought experiment described in two different ways can lead the same people to opposite conclusions. He concludes that their value lies in the shedding light on the ways in which we think and feel (2011, p. 12). Intellectuals benefit from understanding the workings of intellect as much as the mechanic benefits from understanding the workings of the machine.

Since no one knows for sure whether or not a transcendent realm of truths exists, we cannot determine once and for all whether Plato or Aristotle was correct. One thing of which we can be quite certain, however, is that some students are rationalists and others are empiricists. In a classic article, Felder and Silverman synthesize findings from a number of studies to identify contrasting learning styles (Felder, R.M. and L.K. Silverman, 1988, pp. 674–681). They define a student's learning style by the answers to four questions:

- What type of information does the student prefer: sensory (sights, sounds, and physical sensations), or intuitive (memories, ideas, and insights)?
- How is information received: visual (pictures, diagrams, graphs, and demonstrations), or verbal (sounds, written and spoken words, and formulas)?
- How do they process information: actively (through engagement in physical activity or discussion) or reflectively (through introspection)?
- How does the student progress toward understanding: sequentially (in a logical progression of small incremental steps), or globally (in large jumps, absorbing material randomly)? (Wirz, 2004, p. 2)

The dimensions of this model are a matter of degree and a student's preference for the different styles may change with time or from one subject to another. Yet the data clearly suggests that some students learn best "intuitively" from formulas and principles while others learn best "sensorily" from hands-on experiences or concrete examples. These are the rationalist and empiricist, respectively.

Regardless of whether or not Brown is correct to posit a transcendent world, he is certainly correct to suggest that thought experiments focus mental perception in a productive way for those who are rationalistically inclined. Likewise, Norton is correct to suggest that, for those who are empirically inclined, thought experiments can illustrate an argument in a uniquely compelling way. We can set aside the ancient metaphysical debate over transcendent truth while agreeing that rationalist and empiricist approaches to learning are equally important. The fundamental value of thought experimenting is to reveal these divergent approaches at work.

The interdisciplinary nature of thought experiments

Interdisciplinary is the combination of academic disciplines or schools of thought to produce new perspectives and solutions (Augsburg, 2005). In addition to combining philosophy, history, literature, drama, science, and theology at a surface level, at a deeper level, thought experimenting combines rationalist and empiricist schools of thought. In this way it reminds us why all our disciplines are called "arts and sciences" in the first place: the place where intuition and observation meet.

As we've seen, Lucretius provides us with a simple scientific thought experiment. Its interdisciplinarity stems primarily from its invitation to both rationalist and empiricist analysis. Judith Jarvis Thomson provides a famous

example of a much more complicated ethics thought experiment that is interdisciplinary on another level (1971, pp. 47–8).

You wake up in the hospital to find a famous violinist dependent on you for life support. He is attached to you through various tubes and will need you to remain in bed next to him for nine months. Exactly how and why this happened can be elaborated in a number of different ways. For our purposes, suffice it to say that when I present it to the students I tell them that at the end of the class period we are going to take a vote: will you sacrifice a significant portion of your life to save this stranger or will you pull the plug?

The student reactions are interesting. Here is a sampling:

> "I'm going to sue the paramedic who did this to me without my permission!"
> "But if you pull the plug, you kill him. Killing an innocent person is murder, which violates the sixth commandment."
> "Killing isn't the same as letting him die. I have the right to my body; he doesn't have a right to it."

In these three reactions, we see a legalistic thinker, a theological thinker, and a political thinker hashing it out. Thomson used the thought experiment as a model for thinking about abortion. This adds a literary dimension: does the violinist scenario provide an apt comparison to an unwanted pregnancy? Students have a lot to say about that too. Their writing assignment for this week will be to take a position on the question of whether or not abortion is ever morally permissible. The students always express a wide variety of views on this issue because they know, by observing the neutral role I adopt in facilitating the thought experiment, that I will not judge them according to what position they take, but rather according to how well they argue for it.

Back in my lecturing days I always had trouble explaining to my students how the theory of evolution challenges the argument from design, according to which, as Thomas Aquinas famously argues, God must exist because only he could have created such an extraordinarily complex system of nature (1996–1997, Part 1, Article 3, Question 2). Now I get the point across with the help of a thought experiment.

I tell the students that the CDC has learned that they and everyone in their generation is barren. The human race will soon be extinct. However, we have discovered that the apelike species from which we evolved is still alive in Africa. If we put some members of this species into an "evolution accelerator" we can evolve a new race of humans before our race dies out. Again, they are required to vote: should we do it?

A sample of common student reactions:

> "How could humans create other humans from animals? They would lack the divine spark that makes us different from the animals."
> "How could evolution be accomplished in such a short time? It would take millions of years and the random mutations might lead to some creature other than human."
> "How would the new race interact with our old race? Would we intermix or keep them separate and would we tell them that we made them?"

In these reactions, a theological thinker, a scientific thinker and a sociological thinker pushes the limits of their understanding. Their writing assignment for the week will not directly concern this thought experiment. It will ask them to take a position on whether or not evolution defeats the design argument for the existence of God. In my experience, not only do the students have more fun in class, imagining and laughing about strange possibilities, they also produce higher quality papers.

Finally I will mention one of my favorite class thought experiments: what if you alone exist and all of reality is, as René Descartes suggested, an illusion imposed on you by an evil genius? (1993, bk. I) This is a deep, purely philosophical meditation. And yet it is perennially popular. It takes a while to explain to the students why, under these circumstances, they cannot know whether the world exists, or whether they have a body, or whether any of their memories are true. Once they understand the dire nature of their situation, however, they spontaneously recreate Descartes' moment of eureka, each in their own way. At the hands of this evil genius, I ask them, what can you be certain of?

> "That I exist."
> "That I am perceiving something."
> "That I am thinking."

Here we have a barebones confrontation between an ontological thinker, an empiricist, and a rationalist. They are each right within their own systems.

When teaching through thought experiments, the instructor is forced to refrain from imposing a hidden agenda on the discussion. By exploring a problematic scenario and then being required to vote on its resolution, students discover how they think in contrast to how others think; they have to decide for themselves the best way to the truth. Hence this pedagogy is a

propaedeutic to professor proselytizing in the university. As Oskar Gruenwald, editor of the Journal of Interdisciplinary Studies, argues, "the university needs to re-dedicate itself to the search for truth about ourselves and the world without cant and politically correct ideologies" (2011, p. 16). Thought experimenting is a specific proposal for how to accomplish this in the classroom.

The challenge of interdisciplinary pedagogy

So far I have argued that thought experiments promote interdisciplinary classroom discussions and that these are valuable in university education because they fostering each student's individual approach to the search for truth.

Nevertheless, interdisciplinary pedagogy is challenging due to its complexity. As Harvard education researcher Zachary Stein et al. argue,

> Interdisciplinary syntheses are among the most epistemologically complex endeavours that humans can attempt. This complexity arises primarily from the deep differences of perspective that must be bridged in order to carry out interdisciplinary projects. That is, different methods and disciplines frame different perspectives and thus generate different kinds of knowledge (2008, p. 402).

Stein et al. report the results of an experiment aimed at gaging the value of interdisciplinary discussions among professionals. In one such experiment, an accomplished mathematician and an accomplished neuroscientist were asked to discuss problems at the intersection of their fields. Analysing the transcriptions, Stein et al. write,

> These conversations were attempts to advance knowledge by bringing together and synthesising diverse and sophisticated perspectives on issues of great importance (from mathematics to morality and from physics to politics). But instead of reading like constructive dialogues, these conversations often read like a set of juxtaposed monologues. In both cases the two experts find it difficult to avoid privileging the methodological perspectives they hold dear. And all too often the result is disciplinary ships passing in the epistemological night. (2008, p. 405)

Stein et al. conclude that the experiments demonstrate that interdisciplinary studies are prone to two problems that stem from differences between levels

of analysis and differences between basic viewpoints. Let us look at each in turn in connection with a course in the bioethics of learning disabilities.

The levels of analysis problem arises when discussants are using different explanatory frameworks. One discussant may be interested in understanding how various pathologies are diagnosed and treated, while another is interested in understanding how they are accommodated within the school system. Although each discussant uses the term "ADHD" accurately, one has a biological understanding of it; the other, institutional.

The basic viewpoints problem arises when discussants start with incompatible assumptions about the world. One discussant may believe that the term "ADHD" names a genuinely debilitating physiological disorder while another believes it to be a largely imagined psychosomatic condition. Although the discussants may agree that school systems need to offer accommodations for students with ADHD, they may strongly disagree about their extent.

Stein et al. call for a commitment to methodological pluralism as a means of addressing both of these problems. By methodological pluralism they seem to imply that instructors of interdisciplinary courses should avoid privileging any one methodology over another. They should explicitly respect and call attention to the different levels of analysis and different basic viewpoints without attempting to reduce them to a single approach.

In my view, the problem with this proposal is that a professor trained in one field is really in no position to instruct students in areas outside of that area of specialization. How is a historian to evaluate the kind of knowledge generated by a budding psychologist? Will she be able to distinguish good psychological methodology from bad? If not, then methodological pluralism just opens the door to anything goes—a "free-for-all" gab session with no educational value whatsoever.

While I fully endorse Stein et al.'s insistence on respecting and calling attention to differences, I think there is a sense in which different approaches must be reduced or at least subordinated to a single approach within a course. Although my class discussions are interdisciplinary, my class is still a philosophy class. This means that the papers the students write for their final grades are philosophy papers. I am trained in philosophy. I am not competent to judge a literature paper or a psychology paper. Hence the literary and psychological thinkers in my class will have to learn how to incorporate their insights into philosophical discourse. I would expect the same subordination to occur in any interdisciplinary course. The very term "interdisciplinary," after all, presupposes the underlying presence of the disciplines. The disciplines demarcate powerful methodologies that establish criteria of

excellence. Although these methodologies change and grow slowly over time, it is still up to instructors who have mastered these methodologies to pass them on to the next generation.

Hence it seems that, as Jennifer Jesse, co-editor of the American Journal of Theology and Philosophy, argues, one cannot be interdisciplinary without being self-consciously so (2011, p. 72). In fact, being interdisciplinary largely amounts to introducing a metanarrative into class discussions that constantly highlights the plurality of our thinking with the aim of ultimately unifying us in the age-old quest for truth.

Conclusion

We may agree that lecturing creates a perniciously passive classroom, but how to create a lively and meaningful discussion? Playing the usual "I ask and you answer" game creates a predictable and inauthentic exchange. Thought experiment, in contrast, is spontaneous, mutually insightful for teacher and student, and fun. Thought experiment is the sine qua non of philosophy; catalogues of famous and not so famous ones can readily be found (Schick, 2013; Tittle, 2004). But remember: philosophy is the mother of all the disciplines. To this extent philosophy is a welcome complement to any university course, from history (De Mey, 2003), to economics (Stringham, 2008), to math (Clegg, 2003, pp. 239–242).

University education should aim to produce philosophical historians, philosophical economists, and philosophical mathematicians. What is the distinguishing mark of the philosopher? According to Edouard Machery, Sorbonne educated Resident Fellow of the Center for Philosophy of Science at the University of Pittsburgh,

> philosophers are less likely to blindly accept their intuitions and more likely to submit those intuitions to scrutiny. Philosophers ponder; they question what spontaneously seems to be the case; they readily take a skeptical eye toward how things seem to them (2011, p. 211).

By thought experimenting about what is possible, philosophers learn to question what is allegedly actual. By reflecting on their own and others' thought processes, they learn to trust the process of inquiry rather than authority.

How can we ensure that the university will continue to be a source of knowledge and inspiration for the next generation and into the future? Oskar Gruenwald argues, "Philosophy can help here in suggesting not only the obvious distinctions concerning appropriate methodologies in the natural

sciences, social sciences, and humanities, but also concerning the need for more global, interdisciplinary approaches for greater understanding" (1999, p. 163). As the number of disciplines continues to multiply and the interconnections among them become increasingly complicated, the university must stay rooted in its philosophical past and thought experiment is a promising way to accomplish this.

References

Aristotle, De Anima (On the Soul). (1986). Hugh Lawson-Tancred, (tr.), New York, NY: Penguin Classics.

Augsburg, T. (2005). Becoming interdisciplinary: An introduction to interdisciplinary studies. Dubuque, IA: Kendall/Hunt.

Bailey, Cyril. (1950). Lucretius on the Nature of Things, (translation of De Rerum Naturae), ninth reprint, Oxford, UK: Clarendon Press.

Brendel, Elke. (2004). Intuition Pumps and the Proper Use of Thought Experiments. Dialectica 58(1), 89–108.

Brooks, David H. M. (1994). The Method of Thought Experiment. Metaphilosophy, 25: 71–83.

Brown, James R. (2004). Why Thought Experiments Do Transcend Empiricism. in C. Hitchcock (ed.), Contemporary Debates in the Philosophy of Science, Malden, MA: Blackwell, 23–43.

——, 2010, Laboratory of the Mind: Thought Experiments in the Natural Sciences, [2nd edition], London, UK: Routledge Second Edition.

Clatterbuck, Hayley. (2013). The epistemology of thought experiments: A non-eliminativist, non-platonic account. European Journal for Philosophy of Science, 3: 309–329.

Clegg, Brian. (2003). Infinity: The Quest to Think the Unthinkable, London, U.K.:Constable & Robinson Ltd.

Cooper, Rachel. (2005). Thought Experiments. Metaphilosophy, 36(3), 328–347.

De Mey, Tim and Erik Weber. (2003). Explanations and Thought Experiment in History. History and Theory 42(1), 28–38.

Descartes, René. (1993). Meditations on First Philosophy, Stanley Tweyman (ed.), London, U.K.: Routledge.

Elgin, Catherine Z. (1993). Understanding: Art and Science. Synthese, 95: 13–28.

—— (2007). The Laboratory of the Mind. in J. Gibson (ed.), A Sense of the World: Essays on Fiction, Narrative and Knowledge, Oxon: Routledge-Taylor Francis, 43–54.

—— (2014). Fiction as Thought Experiment. Perspectives on Science, 22: 221–241.
—— (2007). Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium. Midwest Studies in Philosophy of Science, 31: 68–89.
Felder, R.M. and L.K. Silverman. (1988). Learning and Teaching Styles in Engineering Education. Engr. Education, 78(7), 674–681.
Goodenough, Jeremy. (2011). The Trouble with Thought Experiments. Theoretical and Applied Ethics, 1(2), 7–12.
Gruenwald, Oskar. (1999). Philosophy as Creative Discovery: Science, Ethics, and Faith. Journal of Interdisciplinary Studies XI: 157–174.
—— (2011). The University as Quest for Truth. Journal of Interdisciplinary Studies XXIII: 1–18.
Horowitz, Tamara, and Massey, Gerald (eds.) (1991). Thought Experiments in Science and Philosophy, Lanham: Rowman & Littlefield.
Jesse, Jennifer G. (2011). Reflections on the Benefits and Risks of Interdisciplinary Study in Theology, Philosophy, and Literature. American Journal of Theology and Philosophy, 32(1), 62–73.
Machery, Edouard. (2011). Thought Experiments and Philosophical Knowledge. Metaphilosophy 42(3), 192–214.
Meinong, Alexius. (1907). Das Gedankenexperiment. in Über die Stellung der Gegenstandstheorie im System der Wissenschaften (reprinted in the fifth volume of the collected works of Meinong, edited by Rudolf Haller and Rudolf Kindinger, Graz-Austria: Akademische Druck und Verlagsanstalt, 1973, 273–283 [67–77]).
Meynell, Letitia. (2014). Imagination and insight: a new acount of the content of thought experiments. Synthese 191:4149–4168.
Norton, John D. (2004a). On Thought Experiments: Is There More to the Argument? Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association, Philosophy of Science, 71: 1139–1151.
—— (2004b). Why Thought Experiments Do Not Transcend Empiricism. In C. Hitchcock (ed.), Contemporary Debates in the Philosophy of Science, Oxford, UK: Blackwell, 44–66.
Perler, Dominik. (2008). Thought Experiments: The Methodological Function of Angels in Late Medieval Epistemology. in Isabel Iribarren and Markus Lenz (eds.), Angels in Medieval Philosophical Inquiry, Aldershot: Ashgate, 143–153.
Plato. (2002a). Crito. in G.M.A. Grube (tr.), Five Dialogues, Indianapolis, Indiana: Hackett, 45–57.
—— (2002b). Meno. in G.M.A. Grube (tr.), Five Dialogues, Indianapolis, Indiana: Hackett, 58–93.

Rescher, Nicholas. (2005). What If?: Thought Experimentation in Philosophy, New Brunswick, NJ: Transaction Publishers.

Schick, Theodore and Lewis Vaughn, (eds.) (2013). Doing Philosophy: An Introduction Through Thought Experiments [5th edition], Boston, MA: McGraw Hill Higher Education.

Stein, Zachary, Michael Connell, and Howard Gardner. (2008). Exercising Quality Control in Interdisciplinary Education: Toward an Epistemologically Responsible Approach. Journal of Philosophy of Education, 42(3–4), 401–414.

Stringham, Edward P. and Nicholas A. Snow. (2008). The broken trailer fallacy: Seeing the unseen effects of government policies in post-Katrina New Orleans. International Journal of Social Economics 35(7): 480–9.

Thomas Aquinas, Summa Theologica. (1996–1997). Fathers of the English Dominican Province (trs.), New Advent Inc.

Thomson, Judith Jarvis. (1971). A Defense of Abortion. Philosophy & Public Affairs, 1(1), 47–66.

Tittle, Peg. (2004). What if….Collected Thought Experiments in Philosophy, Upper Saddle River, NJ: Pearson.

Weber, Erik and Tim DeMey. (2003). Explanation and Thought Experiments in History. History and Theory, 42: 28–38.

Wirz, Dick. (2004). Students' Learning Styles vs. Professors' Teaching Styles. Inquiry, 9(1), pp. 1–5.

Sharon Kaye
John Carroll University
skaye@jcu.edu

On Fischer and Frankfurt-style Cases

William Simkulet

Abstract
Almost everyone believes that moral responsibility requires control; however, philosophers disagree about whether this control is compatible with universal causal determinism. Many philosophers argue that it is not, and to illustrate this intuition they turn to the principle of alternate possibilities (PAP) to demonstrate this incompatibility. According to PAP, for an agent to be morally responsible for her action, she must have been able to do otherwise. If our actions are causally necessitated by circumstances that occurred long before we were ever born, it wouldn't make sense to say we are responsible because we lack both alternate possibilities and control. Recently, compatibilists – starting with Harry Frankfurt – have attempted to construct counter-examples to PAP – Frankfurt-style cases – in which an agent is said to be morally responsible while lacking alternate possibilities. In "The Frankfurt Cases: The Moral of the Stories," John Martin Fischer defends Frankfurt-style cases from what he calls "The Dilemma Defense." Here I argue Fischer's defense fails.

On Fischer and Frankfurt-style Cases
Although most philosophers believe moral responsibility requires control, they disagree about what kind of control is sufficient for moral responsibility. Notably, there is disagreement between as to whether moral responsibility is compatible with universal causal determinism, where *universal causal determinism* is the theory that there is only one possible future and that future is completely causally necessitated by the laws of nature and circumstances of the distant past. *Compatibilists* believe moral responsibility is compatible with determinism, while *incompatibilists* argue that responsibility is fundamentally incompatible with determinism. To circumvent complex metaphysical and metaethical issues, many incompatibilists turn to the principle of alternate possibilities to illustrate this incompatibility:

Filosofiska Notiser, Årgång 5, Nr. 1, Maj 2018, 19–40

> *Principle of Alternate Possibilities* (PAP) – A necessary (but not sufficient) condition for agent *a*'s being morally responsible for *x* is that *a* could have done other than *x*.

Traditionally, both compatibilists and incompatibilists have found PAP to be a commonsense, intuitively plausible moral principle, yet incompatibilists argue that if PAP is true, moral responsibility is incompatible with universal causal determinism because if universal causal determinism is true, then there is only one possible future and no one can actually do other than they actually do – they lack genuine alternate possibilities.

In "Alternate Possibilities and Moral Responsibility," Harry Frankfurt constructs a case intended to be an open-ended counter-example to PAP in which an agent – Jones – is said to (i) *uncontroversially* lack alternate possibilities and yet (ii) be *uncontroversially* morally responsible for his actions. To avoid begging the question about the compatibility of determinism and responsibility, Frankfurt intends the case to be *metaphysically neutral* such that regardless of one's metaphysical beliefs about causation, one would be inclined to believe Jones is uncontroversially responsible despite lacking alternate possibilities.

Critics argue that Frankfurt-style cases are not true counter-examples to PAP because they cannot be constructed in such a way as to show both (i) and (ii) – either Jones (~i) has alternate possibilities, or he (~ii) is not uncontroversially morally responsible; call this *the dilemma defense.*[1] In "The Frankfurt Cases: The Moral of the Stories," John Martin Fischer sets out to defend Frankfurt-style cases from the dilemma defense, and attempts to show that (i) and (ii) can be true of such a case by building upon recent work in the field.[2] Here I will look at Frankfurt's original case, the dilemma defense, and Fischer's response, and argue that Fischer's response is unsatisfying because it abandons the open-endedness of Frankfurt-style cases. I will end by discussing what a satisfactory Frankfurt-style case would look like, and argue that the continuing inability of compatibilists to construct such a case suggests their theory is inconsistent with our metaethical intuitions about moral responsibility.

[1] See Robert Kane (1985, 1996); David Widerker (1995), Carl Ginet (1996); Stewart Goetz (2005).
[2] See Derk Pereboom (2008).

I. Frankfurt on PAP

In "Alternate Possibilities and Moral Responsibility," Harry Frankfurt makes two substantive arguments against PAP:

(1) PAP derives its intuitive force from the coercion principle, the coercion principle is false, and thus PAP lacks intuitive force.

(2) It is possible to construct a case in which an agent (i) uncontroversially lacks alternate possibilities, and yet (ii) is uncontroversially blameworthy, so PAP is false.

The latter argument has been the focus of much recent work in philosophy, while the first has failed to garner much attention. Here I will look at both.

Frankfurt's first argument contends that there is an intuitive connection between PAP and the coercion principle:

> *Coercion Principle* (CP) – A necessary (but not sufficient) condition for agent *a*'s being morally responsible for *x* is that *a*'s action *x* is uncoerced.

One might appeal to a principle like CP to explain why it doesn't make sense to hold coerced agents morally responsible in cases like this:

> *Smith*[1]: White kidnapped bank manager Smith[1]'s family and told him that he will execute them in an hour unless he brings him a million dollars from the bank without alerting the police. Smith[1] believes White's threat, loves his family, steals the money, and gives it to White.

Many people do not think it is appropriate to blame or punish Smith[1] for stealing the money under these circumstances. CP offers an explanation why – Smith[1] was coerced, and coercion undermines moral responsibility. But CP is false.

> *Smith*[2]: White kidnapped bank manager Smith[2]'s cilantro plant and told him that he will set it on fire in an hour unless he brings him a million dollars from the bank without alerting the police. Smith[1] believes White's threat, likes cilantro, steals the money, and gives it to White.

Even if Smith[2] was coerced by White (perhaps due to an atypical attachment to his cilantro plant), it doesn't make sense to say that he isn't responsible for what he does, and thus Smith[2] is a counter-example to CP. Smith[2] is uncontroversially blameworthy.

Furthermore, it makes sense to say that Smith[1] *is* responsible for his action – he took quite a risk to save his family's life and this is *prima facie* praiseworthy. James Rachels argues the right thing to do is the thing that one has the best reasons to do, and it is at least plausible to say that Smith[1] has more reasons to take the money than not to.[3] It seems coerced agents can be either praiseworthy or blameworthy depending upon the reasons they act upon.

Frankfurt says:

> Now the doctrine that coercion and moral responsibility are mutually exclusive may appear to be no more than a somewhat particularized version of [PAP]. It is natural enough to say of a person who has been coerced to something that he could not have done otherwise. And it may easily seem that being coerced deprives a person of freedom and of moral responsibility simply because it is a special case of being unable to do otherwise. The principle of alternate possibilities may in this way derive some credibility from its association with the very plausible proposition that moral responsibility is excluded by coercion (1969, 830–831).

It's not clear why Frankfurt believes the coercion principle is merely a particularized version of PAP as it seems coerced agents often can do otherwise. However, he does suggest that if one has been sufficiently coerced, then one lacks alternate possibilities because the alternatives are too horrific. This is quite suspect – surely people freely choose to act in horrific ways, so the mere fact that an alternate possibility is horrific does not preclude it from being a possibility (although it might make sense to say that choosing such a possible action would be immoral or irrational).

It is possible that an agent can be psychologically constituted such that certain scenarios trigger what Eddy Nahmias calls *bypassing*, the circumventing of an agent's normal moral deliberation processes in such a way that is inconsistent with moral responsibility. Nahmias's position is consistent with Frankfurt's – the existence of deterministic bypassing invites

[3] See Rachels (2003).

confusion, potentially leading interpreters to the conclusion that determinism is responsibility-undermining, rather than bypassing. Note that for Nahmias, an agent is morally responsible for the choices that result from their normal moral deliberation, so if $Smith^1$ and $Smith^2$ were bypassed, neither would be responsible on his view.

To illustrate the falsity of CP, and undermine the intuitive force behind PAP, Frankfurt constructs a series of three cases:

1. Jones freely chooses to perform action A, and freely does so. As it so happens, a would-be coercer requests that he do A and threatens a harsh penalty for refusing. However, Jones ignored this threat and did A for his own reasons.
2. Jones freely chooses to perform action A. However, before he does A, a coercer requests that he do A and threatens a harsh penalty for refusing. Jones is stampeded, forgets his original free choice, and does as the coercer asks to avoid the penalty.
3. Jones freely chooses to perform action A. However, before he does A, someone requests that he do A and threatens a harsh penalty for refusing. Jones does not ignore the threat, and had he not already freely chosen to A, he would do A to avoid the harsh penalty. However, he freely As for his own reasons.

Frankfurt contends that at least one of these cases is a counter-example to CP, and by demonstrating the falsity of the coercion principle, Frankfurt believes he has undermined the appeal of PAP (834).

Anticipating criticism, Frankfurt constructs a final case designed to appease even the most ardent skeptics. In this case, he turns his attention away from CP towards PAP itself and attempts to construct a case in which an agent (i) uncontroversially lacks alternate possibilities, yet (ii) is uncontroversially morally responsible for his action. A concise version follows:

4. Black wants Jones to perform action A, and has a means of forcing Jones to do A. However, he would rather Jones freely choose to do A. Black has observed that in the past, whenever Jones chooses between an A and ~A, he consistently twitches before the choosing ~A. As it so happens, this time Jones doesn't twitch and freely chooses to A. Black doesn't intervene.

Frankfurt contends that in this case:

(i) Uncontroversially, Jones could not have done otherwise. (Jones lacks alternate possibilities.)

(ii) Jones is uncontroversially morally responsible for his action.

If (i) and (ii) are true, this is a persuasive counter-example to PAP. For Frankfurt, it is not enough to argue that it makes sense to say Jones is responsible despite being unable to do otherwise; rather he means to show that this is *uncontroversially* the case. His goal is to construct a persuasive counter-example to PAP – one that will convince anyone, regardless of their metaethical intuitions regarding the compatibility of free will and determinism, that PAP is false.

Indeed, many find this case convincing. In "Incompatibilism and the Avoidability of Blame", Michael Otsuka, convinced of Frankfurt's success, proposes an alternative to PAP, the *principle of avoidable blame* (PAB), according to which for one to be morally responsible, one must have had the possibility to act in a manner in which one would have been entirely morally blameless. Otsuka's principle strikes me as a plausible replacement for the coercion principle, but not PAP. PAB explains our intuitions in traditional coercion cases, but requires the same kind of alternate possibilities as PAP. Otsuka believes that Jones is wholly morally culpable for his action if he chooses it freely, but that he would have been entirely morally blameless had Black forced him to act. But this just is to say that Otsuka believes that Jones has alternate possibilities – one possibility in which he freely chooses to A, and one in which Black compels him to choose A. Jones has alternate possibilities, what he lacks are *alternate outcomes*; this is to say that Frankfurt's case suggests Jones can freely A or be forced to A by Black. In either possibility, the outcome is the same – Jones As; what differs is how Jones As (freely or by compulsion) and how responsible he is for it (responsible or not responsible).

II. Fischer on the Dilemma Defense

In "The Frankfurt Cases: The Moral of the Stories", John Martin Fischer contends the primary threat to the legitimacy of Frankfurt cases comes from the *dilemma defense*. According to this argument, either:

(~i) Jones could have done otherwise.

or

(~ii) Jones is not uncontroversially morally responsible.

The argument goes as follows – either Jones's actions are (a) undetermined or (b) determined. If (a), then Black cannot know what he is going to do before he does it, and thus Black cannot cut off alternate possibilities. If (b), then Jones is not uncontroversially morally responsible because there is

substantial controversy over whether determinism is compatible with moral responsibility. Fischer calls (a) the *indeterministic horn* of the argument, and (b) the *deterministic horn*, and offers an argument against each. I briefly describe each horn, then discuss Fischer's reply.

A. The Indeterministic Horn

Early criticism of Frankfurt cases focused on critiquing the method by which Black is said to know what Jones will do.[4] Frankfurt proposes that Black has discovered that Jones has a tell – a twitch that indicates how he will act in the future. Critics contend this account fails to show both (i) and (ii); either ($\sim i_1$) the twitch occurs after Jones's choice, and thus Black is too late to cut off alternate possibilities, ($\sim i_2$) the twitch occurs before Jones's choice, but is not causally connected to his choice, and thus Black cannot be sure he's cut off alternate possibilities (despite Black's prior observations, Jones can choose to ~A without twitching), or (~ii) the twitch is causally connected to Jones's choice, such that his choice is wholly causally determined either by the twitch, or by a preceding states of affairs, and thus Frankfurt can't help himself to the conclusion that Jones is uncontroversially morally responsible for his actions.

In "Rescuing Frankfurt-Style Cases", Alfred Mele and David Robb introduce a method for circumventing this problem:

> Black wants Bob to steal Ann's car by time *t*. He believes Bob will choose to freely steal the car by that time, but just in case he implants a device in Bob's brain that will causally determine him to steal Ann's car at *t* unless Bob has already freely chosen to steal it. As it so happens, Bob freely chooses to steal the car on his own, and the device doesn't play a causal role.
>
> However, Bob's modified deliberation process is such that were he to freely choose to steal the car at *t*, his free choice is the cause, but if he would be such that he freely chooses to not to steal the car at *t*, his free choice is bypassed by the device.

The key innovation in this case is that the device in question is said to activate at time *t* regardless of what Bob would choose. If Bob would freely choose not to steal the car, or has simply failed to make a choice, the device bypasses Bob's normal deliberation method and forces him to choose to steal

[4] See Robert Kane (1985, 1996); David Widerker (1995).

the car. If, however, Bob would freely choose to steal the car, although the device activates, it has no effect, and Bob chooses to steal the car of his own free will. The primary problem with this version of the case is that it's not clear that a device could be constructed in such a way that it would be ineffective when Bob freely chooses what Black wants, but effective when Bob would choose otherwise. In a sense, Mele and Robb want it both ways – Bob acts freely when he does what Black wants, and is bypassed before his choice when he counterfactually would have acted otherwise. The reason why Mele and Robb's device is said to be able to do this is because it activates during Bob's deliberation process, and they offer a mechanistic account of Bob's deliberation process such that the device is the cause if and only if Bob would counterfactually have freely chosen otherwise.

Both Fischer and Derk Pereboom construct Frankfurt-style cases where similar devices play an important role[5], but fail to adopt Mele and Robb's account of how the device works, thus leaving their position open to the questions about how the device knows when to intervene.

Rather than adopt Mele and Robb's mechanistic causal story of agency, Fischer turns to the work of David Hunt and Derk Pereboom, in which they contend that the relevant question is not whether the agent in a Frankfurt-style case has alternate possibilities, but whether she has robust alternate possibilities.[6] Pereboom defines *robustness* in the following way:

> For an alternative possibility to be relevant per se to explaining an agent's moral responsibility for an action it must satisfy the following characterization: she could have willed something other than what she actually willed such that she understood that by willing it she would thereby have been precluded from the moral responsibility she actually has for the action. (2008, 4–5)

Pereboom's account builds upon Otsuka's PAB – requiring a difference in blameworthiness between options. However, this account is deficient. To have robust alternate possibilities on this account, one needs to have the option to act in a manner in which one would lack moral responsibility of any kind – but this is absurd! Moral agents are morally responsible for their free choices, blameworthy for their vicious ones, and praiseworthy – or at least blameless – for others. This is to say that when a moral agent acts as a moral

[5] See Fischer (2010), Pereboom (2008).
[6] See Hunt (2000), Pereboom (2001).

agent, even when she chooses to act in a way in which is neither praiseworthy nor blameworthy (for example, choosing between two morally equivalent options), she is still morally responsible for her choice. Otsuka's PAB requires the ability to act in such a way in which one is entirely blameless, not entirely lacking moral responsibility. Such an agent may be praiseworthy (as with Smith[1]) or responsible but neither praiseworthy nor blameworthy. In "On Robust Alternate Possibilities and the Tax Evasion Case," I propose an account of robust alternate possibilities that includes Pereboom's awareness criteria, but more closely resembles Otsuka's account:

> An agent has a robust alternate possibility if and only if she could have acted in a manner in which she believes [or would believe] she would have been differently morally responsible (2015a, 103).

On my view, for each action one is morally responsible for, one has at least one robust alternate possibility in which, if one is blameworthy, one would be entirely morally blameless, and if one is praiseworthy, one would be entirely morally praiseless.[7]

In Frankfurt-style cases, we have good reason to think the agent in question has robust alternate possibilities. In Frankfurt's original case it makes sense to say that Jones believes he has multiple possibilities – that he can freely choose to A or ~A. Suppose Jones believes freely choosing A is prima facie more morally blameworthy than freely choosing ~A. If Jones were to freely choose A, sans Black's device, his choosing ~A would uncontroversially be a robust alternate possibility. But suppose Jones chooses ~A believing it to be morally superior to A, and Black intervenes after this choice forcing him to A against his will. His free choice to ~A is praiseworthy, if only because Jones believes it would lead him to ~Aing; the fact that Black's machinations prevent him from doing what he chooses is irrelevant. When Jones As in such a case, he does so against his will and it doesn't make sense to hold him either blameworthy or praiseworthy as his actions are outside of his control.

If Black intervenes *before* Jones's choice, then Jones is not responsible because he doesn't choose anything; Black has bypassed his free will. But let us suppose Black uses a device like the one imagined by Mele and Robb – a device that intervenes *during* the choice, one that changes the freedom of his

[7] When one acts in a blameworthy fashion, one's robust alternate possibility might be such that she would be not only entirely morally blameless, but also such that she would be praiseworthy.

choice depended on the expected outcome – if Jones would freely A, the device does nothing and Jones's free choice causes his Aing, but if Jones would counterfactually have freely ~Aed, the device causally determined his Aing. Here too, I think, Jones has robust alternate possibilities. By assumption, Jones believes:

(1) He would be blameworthy for freely choosing to A.

(2) He would be differently responsible (less blameworthy, both blameless and praiseless, or praiseworthy) for freely choosing to ~A.

However, surely Jones also believes:

(3) He would lack *any responsibility at all* for the actions causally necessitated by a device that he didn't willingly, knowingly, or negligently trigger.

Belief (3) is not a convenient, *ad hoc* belief – it seems to follow from the commonsense belief that moral responsibility requires control. In Mele and Robb's case, if the device activates Jones would be causally responsible (in some sense) for its activation (as if he had just did as the device would have caused him to do, the device would have remained dormant); however he would have no more moral responsibility for this than a hiker that unintentionally and non-negligently steps in a well-disguised bear-trap along her normal jogging route (well known for its otherwise consistent lack of bear-traps – disguised or otherwise). This is to say that Jones's causal responsibility would be insufficient to hold him morally responsible for the outcome.

Fischer admits traditional Frankfurt cases fail to show that the agent lacks robust alternate possibilities, but he contends that "buffer zone" cases, like ones proposed by Pereboom and Hunt, are promising.[8] In buffer zone cases, it is stipulated that before one is psychologically allowed to consider doing otherwise, they have to meet certain criteria, but some device either prevents them from fully meeting these criteria, or prevents them from acting freely once the criteria is met. A concise version of Pereboom's case goes something like this:

> *Tax Evasion:* Joe believes he can get away with cheating on his taxes, but that to do so would be wrong. Joe's strong desire to advance his own self-interest will causally determine him to cheat on his taxes at t_1 unless he chooses against it for moral reasons. To choose against it for moral reasons, Joe must first raise his moral attentiveness to an

[8] See Pereboom (2001).

> appropriate level through the exercise of his libertarian free will. (Joe cannot act on a whim.) However, even if he were to attain this level of moral attentiveness, his libertarian free will allows him to choose either to pay his taxes or cheat.
>
> Unbeknownst to Joe, a neuroscientist has implanted a device in his brain which is triggered by him reaching the appropriate level of moral attentiveness. If triggered, the device would rob him of his libertarian free will and causally determines him to cheat on his taxes. As it so happens, Joe never uses his libertarian free will to raise his moral attentiveness level "and he chooses to evade taxes on his own, while the device remains idle." (2001, 9–10)

Pereboom contends two things are true in this case:

(i*) Joe (uncontroversially) lacks robust alternate possibilities.

(ii) Joe is uncontroversially morally responsible.

Perhaps the most notable problem with buffer zone cases is that they draw our attention away from the relevant issue – Joe's free choice. To explain how the device can successfully cut off (robust) alternate possibilities, Pereboom has to stipulate that Joe's deliberation process works in a rather specific way – a way that we're told is libertarian/indeterministic, despite the fact Pereboom expressly contends it does not allow for Joe to act on a whim (the ability to act on a whim would constitute a robust alternate possibility, so Joe can't possess it.) If Joe were to act in such a way that, *sans* device, would allow him to choose not to cheat on his taxes, the device would activate, interfering with, and bypassing, Joe's normal deliberation method in such a way that Nahmias would conclude is responsibility-undermining. But don't fret, we're told that as it so happens Joe never bothers to try to do otherwise, he does what he was predisposed to do, and the device does nothing. Thus, Pereboom contends, it makes sense to say Joe freely chooses to cheat on his taxes – he does what he wants to do – and questions about the plausibility of the Joe's deliberation process or the device fall by the wayside, as they don't play a role in his choice.

This is quite a feat of sleight of hand – Joe's choice architecture is said to be indeterministic, despite his inability to act on a whim. Joe's said to have alternate possibilities, but due to the device these alternate possibilities can never change how he will behave. But as it so happens, Joe freely chooses not to act on these (ineffective) alternate possibilities, and he does what he was pre-disposed to do, and despite not exercising his indeterministic free

will, we're told this is his free choice. One could imagine Pereboom saying "Pay no attention to the choice architecture behind the curtain" that is said to cut off robust alternate possibilities, but all this serves to obfuscate the feature of the case that is actually causally effective – how Joe comes to be predisposed to act in his self-interest in the first place.

Suppose that a while back, Black wanted Joe to act in his self-interest and implanted a separate device that bypasses his normal deliberation process and forced Joe to be predisposed to act in his self-interest. If Joe's current predisposition to act self-interestedly is outside of his control in this way, both compatibilists and incompatibilists alike would conclude Joe is blameless for this predisposition. The fact that his (supposedly) libertarian free will would (*sans* neuroscientist's device) give him a second chance to fix this mistake is only somewhat relevant to determining his freedom or responsibility for acting on this predisposition, but in Tax Evasion, Joe would lack even this second chance.

If Joe's predisposition to act self-interestedly was brought about deterministically, incompatibilists would reject (ii). However, if it was brought about in an appropriate indeterministic way, then Pereboom has failed to show that (i*) is true, as Joe might very well have had robust alternate possibilities with regards to setting his predisposition – for example, either he could freely choose to act in such a way that would lead to a predisposition to act self-interestedly or to act in some other way. Complicating the issue, acting in such a way that conditions one to act self-interestedly is not inherently immoral – often the right thing is also the self-interested thing to do – for example, it is often both in your self-interest and morally correct to exercise, eat right, and practice good hygiene. Thus, even if Joe freely acted in such a way that conditioned himself to act in his self-interest, this is not necessarily morally blameworthy.

But now, let us look at a feature of Joe's deliberation process necessary, but not sufficient, to get him to reconsider cheating on his taxes – raising his moral attentiveness level. Assume the following:

1) Joe knows that he can only change his predisposition by raising his moral attentiveness level.
2) Joe knows how to raise his moral attentiveness level.
3) Joe believes that if he does nothing, he will act immorally by cheating on his taxes.
4) Joe believes there are no other, more pressing morally relevant concerns that ought to take precedence.

If these are true, then I think it makes sense to say that Joe *does* have robust alternate possibilities – he can either freely choose to raise his moral attentiveness level, or he can freely neglect to do so. Consider the following principle:

> *Stepping Stone Principle (SSP):* If *a* has a moral obligation to do *x*, and *a* believes *y* is necessary to do *x*, then *a* has a moral obligation to do *y*.

Joe believes that a necessary (although not sufficient) step to avoid cheating on his taxes is to raise his moral attentiveness level, and if SSP or something like it is true, Joe has a moral obligation to raise his moral attentiveness level so that he can avoid cheating on his taxes. Failure to do so is *prima facie* blameworthy. This means that Joe does have robust alternate possibilities in Tax Evasion – he can either raise his moral attentiveness level or he can refrain from doing so. He believes he ought to do the former, yet does the latter, and thus he fails to do what he believes he ought to do. But this *just is* to say that he had robust alternate possibilities – possibilities that, if he acted on them, would lead to his being differently responsible.

In contrast, if we do not assume 1–4, Joe is quite the bizarre agent, inexplicably ignorant about how to make decisions for himself. Such an agent would be incompetent and it wouldn't make sense to hold him blameworthy for cheating on his taxes because he lacked the awareness to do otherwise. At the very least, Joe would not be *uncontroversially* morally responsible for his actions.

Fischer admits he fails to offer a satisfactory defense against the indeterministic horn of the dilemma, but says:

> I simply want to motivate the idea that the jury is still out with respect to the indeterministic horn of the Dilemma Defense. I wish to emphasize that it is not enough to point out that if indeterminism obtains, there will always be *some sort of residual alternative possibility*; the alternative possibility must be of the right sort—it must be sufficiently robust to ground attributions of moral responsibility. Arguably, explicitly indeterministic versions of the Frankfurt cases can be developed in which it is highly plausible that the agent is morally responsible and yet lacks access to robust alternative possibilities. Intuitively, the lack of access to robust alternative

> possibilities in these cases is *irrelevant* to the agent's status as morally responsible (2010, 323).

Fischer's discussion here isn't meant to be definitive. However, it is a substantial attempt to reframe the scope of Frankfurt-style cases; like Pereboom, Fischer wants to move the bar – a successful Frankfurt-style case, he contends here, need only aim to show merely that an agent (uncontroversially) lacks robust alternate possibilities, not just any kind of alternate possibilities.

If an indeterministic Frankfurt-style case can be developed in which an agent (i) uncontroversially lacks any sort of residual alternate possibilities, and yet (ii) is uncontroversially morally responsible, then one would have robbed PAP of its intuitive plausibility, weakening the appeal of incompatibilist stances on moral responsibility. The problem, it seems, is that critics of PAP have not been able to construct such a case.

In absence of such a case, Fischer contends that it should be possible to develop an indeterministic Frankfurt-style case in which an agent (i*) (uncontroversially) lacks robust alternate possibilities, yet (ii) is uncontroversially morally responsible. There are two problems with this approach: (1) It seems critics of PAP have thus far been unable to construct cases of this kind either, and (2) even if critics of PAP are right and any old alternate possibilities are irrelevant to determining an agent's moral responsibility, it is still possible that they correspond with moral agency. Elsewhere, I argue that even residual alternate possibilities may be important indicators of moral responsibility (2015b). Both sides of the free will debate contend moral responsibility requires control, and (at least for incompatibilists), the lack of alternate possibilities is a rather noteworthy indicator that an agent lacks control even if their absence is not the cause of the agent's lack of control.

B. The Deterministic Horn

In an earlier paper "Recent Work on Moral Responsibility," Fischer presented an argument that the deterministic horn of the dilemma defense fails; and offers a refined version of that argument in "The Frankfurt Cases: The Moral of the Stories." He says:

> [S]upposing that we explicitly assume that causal determinism obtains in the cases, it is important first to note that I do not propose that we precipitously conclude, from mere reflection on the cases, that (say)

> Jones is morally responsible for his choice and action. Rather, the initial conclusion is that if he is not morally responsible, it is not because he lacks appropriate alternative possibilities. *This initial conclusion does not beg the question against the incompatibilist* (2010, 323).

Fischer proposes that the compatibilist take this intuition seriously, and examine all the other (non-PAP) reasons that one might conclude that Jones lacks moral responsibility in Frankfurt's case. He continues:

> If such a theorist concludes that, since there are no *other* reasons that constitute good and sufficient reasons to believe that causal determinism rules out moral responsibility, causal determinism is indeed compatible with moral responsibility, this too would *not* beg the question against the incompatibilist (2010, 324).

The primary problem with Fischer's approach is that it doesn't preserve the open-endedness of Frankfurt's argument. The reason PAP has played such a dominant role in the discussion of free will problem is that it allows philosophers to sidestep the complex metaphysical and metaethical disputes at the heart of the free will debate. Frankfurt says that "Practically no one… seems inclined to deny or even question that [PAP] (construed in some way or another) is true" (1969, 829). This is to say that prior to Frankfurt's analysis, compatibilists and incompatibilists largely agreed that PAP (or something like it) was true. This presented a clear problem for the compatibilist position – it was quite an uphill battle to show that alternate possibilities are compatible with universal causal determinism, the theory that there is only one possible future.

The incompatibilists argument is clear and persuasive:

1. PAP – Moral responsibility requires alternate possibilities.
2. If determinism is true, there are no alternate possibilities.
3. Conclusion: Moral responsibility is incompatible with determinism.

I believe the primary virtue of Frankfurt's argument against PAP is that it, too, is intended to circumvent complex metaphysical and metaethical disputes at the heart of the free will debate:

1. Jones (i) uncontroversially lacks alternate possibilities and (ii) is uncontroversially morally responsible.

2. If PAP, either Jones (~i) has alternate possibilities or (~ii) is not morally responsible.
3. Conclusion: PAP is false.

Frankfurt's argument is not an argument for compatibilism, but rather an argument against PAP – if successful, all Frankfurt will have shown is that PAP is false and the (apparent) inconsistency between determinism and alternate possibilities is irrelevant because alternate possibilities are not required for moral responsibility.

The primary virtue of this argument is that it's open-ended – if you're predisposed to believe (i) and (ii), then you are predisposed to believe PAP is false. In contrast, Fischer argues that if you are predisposed to believe (~ii), this isn't enough to show the truth of PAP. But this misses the point – the incompatibilist appealed to PAP simply because most people (compatibilist and incompatibilist, philosopher and layman alike) already believe PAP (or something like it) is true. A successful Frankfurt-style case would immediateely undermine the intuitive plausibility of the moral principle that has allowed incompatibilists to claim their position is *obviously* more consistent with our intuitions than the compatibilist position. Thus, without PAP to bolster the intuitiveness of their position, it is an open question which theory – compatibilism or incompatibilism – is more consistent with our moral intuitions. Of course, compatibilists believe they have a slight edge – our leading scientific theories assume our world acts deterministically, and so the compatibilist position would allow those who assume universal causal determinism to allow for moral responsibility, whereas incompatibilism would lead those who assume universal causal determinism to conclude no one is morally responsible for anything.

Fischer contends that even if you assume determinism and conclude that Jones is not responsible, either (a) you will find there are other responsibility-undermining features of the case independent of the assumption of determinism, or (b) you're begging the question and just assuming determinism itself undermines moral responsibility. Not so.

Regarding (a); compatibilists believe that their view is consistent with most of our commonsense moral intuitions, and thus they believe that upon closer reflection, there is a good chance that one will find other responsibility-undermining features of the case. But this is not the win Fischer thinks it is, as the very fact that arriving at this conclusion requires reflection demonstrates they failed to show (ii). The lack of a Frankfurt-style case that can show both (i) or (i*) and (ii) is quite a problem for the

compatibilist, as if determinism and compatibilism are true, then every instance of genuine moral responsibility is a counter-example to PAP. Yet faced with the threat of the intuitive plausibility of PAP, compatibilists have failed to construct a successful counter-example to PAP. Given the motivation of compatibilists, and their contention that such counter-examples are plentiful, their inability to point to an example that can do both (i) and (ii) is quite foreboding.

Note that concluding (a) is not altogether unproblematic for the proponent of Frankfurt-style cases. If there is some *other* responsibility-undermining feature of a Frankfurt-style case, then it's not a successful Frankfurt-style case (and thus fails to be a counter-example to PAP), but worse still it means that the architect of the case has an unreliable moral intuition – they believe the agent is morally responsible when, in fact, there are responsibility-undermining features of the case that they have overlooked. Not only would such a case fail to show (ii), it would fail to show even (ii*) that the agent is (controversially) morally responsible.

However, Fischer's contention that (b) those that do not find Frankfurt-style cases compelling are question-begging, is more problematic. First Fischer suggests that it is question-begging to conclude that a lack of alternate possibilities (brought about by determinism) is responsibility-undermining. However, before Frankfurt's article, both compatibilists and incompatibilists alike found PAP intuitively plausible and offered distinct accounts of what it required, with incompatibilists arguing PAP required actual alternate possible futures, and compatibilists arguing PAP required something like the ability to do otherwise should one have chosen to – counterfactual alternate possibilities of one kind or another. Thus, if one had the intuition that the agent in a Frankfurt-style case isn't morally responsible, this isn't to say that they're merely begging the question that determinism is incompatible with moral responsibility. After all, many compatibilists – including most that came before Frankfurt – would be happy to explain an agent's lack of moral responsibility in some cases on their lack of (compatibilist) alternate possibilities. For example, Nahmias can sensibly say that when an agent's normal deliberation process is bypassed, say by a neuroscientist's device, she lacks the ability to do otherwise if she had wanted to.

However, there is a deeper problem with Fischer's approach here – it gives up on the open-endedness of Frankfurt's original case. Frankfurt believed that everyone – compatibilist and incompatibilist, philosopher and

layman alike would conclude Jones (i) uncontroversially lacks alternate possibilities, and yet (ii) is uncontroversially morally responsible, undermining the intuitive plausibility of PAP and any intuitive plausibility the incompatibilist position gained by piggybacking on the principle. In contrast, Fischer's contention is that Frankfurt-style cases are plentiful, but that if you doubt the agent in any given Frankfurt-style case is (ii) uncontroversially morally responsible, then (a) don't worry – although it's not a genuine Frankfurt-style case, there is something other than a lack of alternate possibilities undermining moral responsibility, or (b) your intuition is unreliable. While Frankfurt's argument turns on our intuitions, Fischer's argument seems to do the opposite – relying upon intuitions only when they're consistent with compatibilism – in (a) – and dismissing them when they're not – in (b). Thus, it seems Fischer isn't only giving up (i), but (ii); for Fischer, all a counter-example to PAP would have to show is:

An agent:

(i*) lacks robust alternate possibilities.

(ii*) is actually (but possibly controversially) morally responsible (regardless of whether you have this intuition or not).

If Frankfurt is successful, the debate over PAP is over. However, if a Fischer-style Frankfurt-style case could be constructed to show (i*) and (ii*), debate would continue because there is no consensus on what alternate possibilities matter and whether the agent is genuinely morally responsible. This is to say that even if Fischer is right and compatibilism is true, he will have failed to demonstrate this in a persuasive way.

III. Conclusion

Here, I've argued that Frankfurt-style cases are an elegant attempt to reframe the free will debate, circumventing complex philosophical questions about metaethics and metaphysics and resting solely on our commonsense intuitions. If Frankfurt is right, then anyone who reads Jones's case would conclude that he (i) uncontroversially lacks alternate possibilities, and yet (ii) is uncontroversially morally responsible for his actions. Such a case would be a counter-example to PAP, and undermine its intuitive plausibility, in turn undermining the intuitive plausibility of a popular incompatibilist argument against compatibilism.

The primary problem facing proponents of Frankfurt-style cases is that – at least thus far – compatibilists have been unable to construct a case in which they can help themselves to both (i) and (ii). The dilemma defense

argues that either (~i) or (~ii). If Jones's actions are undetermined, then either Black would intervene before Jones's choice, undermining his agency and responsibility such that (~ii), or Black would intervene after Jones's choice, so Jones could have chosen otherwise such that (~i). However, if Jones's actions were determined, anyone that believes determinism undermines moral responsibility would conclude (~ii).

Fischer takes issue with the deterministic horn of the dilemma, contending that it's not fair to conclude that Jones lacks responsibility just because the case assumes determinism. He contends that there may be other reasons why Jones lacks responsibility, or we ought to conclude he is responsible. This is absurd. Our moral intuitions regarding freedom and responsibility are nuanced and complex – the virtue of Frankfurt's strategy is that he intends to circumvent any serious discussion about them by appealing to a case in which he believes everyone – regardless of their moral and metaphysical intuitions – will conclude the agent is not only responsible, but uncontroversially so. If everyone agrees Frankfurt's case is a counter-example to PAP, the debate is over.

As it so happens, Frankfurt's case is not as uncontroversial as intended. Although he intended the case to be metaphysically neutral, the dilemma defense illustrates that he has failed to do so. Fischer contends explicitly indeterministic Frankfurt-style cases could be developed, but refrains from doing so. This is disappointing, as such a case would be sufficient to convince the target audience (incompatibilists) that PAP is unintuitive.

Instead, Fischer focuses on the deterministic horn, challenging the intuitions of those that are unsure of an agent's responsibility in a deterministic Frankfurt-style case. But challenging intuitions in such a case undermines the open-ended, prereflective nature of Frankfurt's argument – in the same way compatibilist Fischer questions the validity of intuitions inconsistent with his metaethical stance on moral responsibility, so, too, could incompatibilists – those that believe responsibility is incompatible with determinism – and hard incompatibilists – those that believe responsibility is incompatible with either determinism or indeterminism – challenge the intuitions of those that agree with Fischer. In other words, rather than circumvent questions about our metaphysical and metaethical beliefs, as Frankfurt and those that appeal to PAP do, Fischer bogs the discussion down in a quagmire of complexity and skepticism regarding (at least some of) our moral intuitions. Perhaps this is as it needs to be – our intuitions regarding metaethics and metaphysics are nuanced and complex and require rigorous

analysis to resolve important questions regarding freedom and responsibility. However, the appeal of Frankfurt-style cases is that they are meant to skip all that.

The literature surrounding Frankfurt-style cases have helped ethicists to make substantial advances in clarifying their position on free will and moral responsibility, and made it easier to articulate important normative principles distinct from PAP, such as Michael Otsuka's principle of avoidable blame. Both compatibilists and incompatibilists contend that their account of moral responsibility is consistent with our moral intuitions. Incompatibilists, it seems, have the upper hand in this debate, owing to the intuitive plausibility of PAP. Frankfurt-style cases represent a means by which compatibilists can challenge the intuitive plausibility of PAP.

If determinism is true, and moral responsibility is compatible with determinism, one would expect successful deterministic Frankfurt-style cases to be quite common and easy to construct. Constructing even a successful indeterministic Frankfurt-style case would be sufficient to undermine the intuitive plausibility of PAP. Yet compatibilists have thus far failed to construct cases of either kind. If our intuitions are compatibilist, the lack of such cases is shocking. However, if our intuitions are incompatibilist, this failure should come as no surprise.

Attempts by Pereboom, Hunt, and Fischer to reframe the debate by focusing on robust alternate possibilities and/or questioning the intuitions of those that don't find the agent in question uncontroversially morally responsible seem defeatist. At best, this muddles the debate in complex philosophical issues, rather than circumvent those issues entirely as Frankfurt intends. Incompatibilists appeal to PAP to circumvent those very issues; if compatibilists hope to show their view is as intuitively plausible as the incompatibilists, they should keep this in mind and, like Frankfurt, attempt to follow suit.

References

Bernard Berofsky, (2012). *Nature's Challenge to Free Will*. Oxford University Press.

Fischer, John Martin and Ravizza, Mark (1998). Morally Responsible People without a Freedom. from R*esponsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge, from *The Determinism and Freedom Philosophy Website*, accessed: 7-05-09 http://www.ucl.ac.uk/~uctytho/dfwCompatFischerRavizza.htm

Fischer, John Martin (1999). Recent Work on Moral Responsibility. *Ethics* 110(1), pp. 93–139.

Fischer, John Martin. (2010). The Frankfurt Cases: The Moral of the Stories. *Philosophical Review* 119(3), pp. 315–336.

Frankfurt, Harry G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy* 66(23), pp. 829–839.

Ginet, Carl. (1996). In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing. *Philosophical Perspectives* 10, pp. 403–417.

Goetz, Stewart (2005). Frankfurt-Style Counter-examples and Begging the Question. *Midwest Studies in Philosophy* 29, pp. 83–105.

Hunt, David (2000). Moral Responsibility and Unavoidable Action. *Philosophical Studies* 97, pp. 195–227.

Kane, Robert (1985). *Free Will and Values*. Albany: State University of New York Press.

Kane, Robert (1996). *The Significance of Free Will.* New York: Oxford University Press.

Kane, Robert (2003). Free Will: New Directions for an Ancient Problem. in Kane (ed.): *Free Will*, Blackwell.

Mele, Alfred R. and Robb, David (1998). Rescuing Frankfurt-Style Cases. *The Philosophical Review*, 107(1), pp. 97–112.

Nahmias, Eddy (2011). "Intuitions about Free Will, Determinism, and Bypassing". in R. Kane, (ed.) *The Oxford Handbook of Free Will 2nd Ed. Oxford University Press*: 555–587

Otsuka, Michael (1998). Incompatibilism and the Avoidability of Blame. *Ethics* 108(4), pp. 685–701.

Pereboom, Derk (2001). *Living Without Free Will*, Cambridge: Cambridge University Press.

Pereboom, Derk (2008). Defending Hard Determinism Again. in *Essays on Free Will and Moral Responsibility*, ed. Nick Trakakis and Daniel Cohen, Newcastle-upon-Tyne: Cambridge Scholars Press, pp. 1–33.

Rachels, James (2003). *The Elements of Moral Philosophy Fourth Edition*, McGraw-Hill.

Simkulet, William (2015a). On Robust Alternate Possibilities and the Tax Evasion Case. *Southwest Philosophy Review* 31(1), pp. 101–107.

Simkulet, William (2015b). On the Signpost Principle of Alternate Possibilities: Why Contemporary Frankfurt-Style Cases are Irrelevant to the Free Will Debate. *Filosofiska Notiser* 2(3), pp 107–120.

Strawson, Galen (1994/2002) The Impossibility of Moral Responsibility. *Philosophical Studies* 75, pp. 5–24. reprinted in *Ethical Theory Classic and Contemporary Readings* by Louis P. Pojman.

Widerker, David (1995). Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities. *Philosophical Review* 104, pp. 247–261.

William Simkulet
Cleveland State University
simkuletwm@yahoo.com

There is no Arrow of Time

Andrea Roselli

Abstract
Instead of the linear temporal description of reality, I illustrate an alternative model which eradicates the concepts of direction and entropy from that of time. Time, intended as a Relationist measure of change, has only the possibility to pass positively or to stay still: the unidimensional mathematical metaphor is misleading, it is not possible to live or experience reality backwards. In light of that, I provide a different reading of the time-reversal invariance of the fundamental laws of physics.

Aim and structure of the paper
In the first paragraph, I claim that the time reversal invariance of local or macroscopic descriptions of physical systems should not be interpreted in a strong ontological sense – we should pay attention to get the right moral from the mathematical translation of a physical situation. In the second, I show the difficulties that emerge when we try to tie together entropy and time: besides the problem of making sense of the expression "entropy of the universe", many convincing thought-experiments could be generated to show how entropy and time are not related. In the third, I maintain that the Second Law of thermodynamics simply states a statistical truth – it doesn't impose a direction on nature. In the fourth and final paragraph, I sketch a Relationist temporal account. Without change (if everything freezes at 0 Kelvin, for example) there is no passage of time. There is no need of an arrow of time, because there are not two different directions, the only possibilities are to change or not to change. Finally, I defend the view from possible charges of circularity, and indicate some possible further developments.

1. The *time-reversal invariance* of the fundamental laws of physics

A very reasonable way to conduct an ontological analysis of our universe is to look at the best scientific theories available, with the goal of determining what they imply about reality. What is not always clear, though, at a time in which mathematics and physics have developed their own language, their

Filosofiska Notiser, Årgång 5, Nr. 1, Maj 2018, 41–56

own practice, is how and to what extent it is possible to translate physical-mathematical models into a definite description of reality. Fundamental science has mainly to do with concepts and structures that are well defined under a mathematical point of view, but undergo some difficulties when it comes to represent those contents intuitively (in the world out there).

A possible reaction consists, of course, in attributing every sort of problem to our intuition. Since we don't experience the fundamental physical reality, the reasoning goes, it is obvious to find its concepts and descriptions exotic. A different possibility, however, is to argue that mathematics and physics have not only their own language, but also their own rules, part – in a certain sense – of their own world. Many times the translation from one world to the other is simple, even obvious, some other times, though, it is not. A classical example is the notion of line – what does it mean that a line has an infinity of points? This has a definite mathematical meaning, but what about its physical meaning?[1] The current scientific formalization of many basic entities (universe, time, space, …) contain a great number of internal rules that are not directly translatable into the physical world that surrounds us. The problem of the so-called arrow of time, I believe, has something to do with it.

The usual approach, when we speak about the problem of the arrow, starts off by acknowledging that the fundamental physical laws are time-reversal invariant (it seems that micro-physics recognizes no directionality of time). Anything that happens could, under a microphysical point of view, happen in reverse. Our experience, on the other hand, seems to tell a very different story. The cause-effect relation is always well aligned past-to-future, and the idea of effects occurring before their causes is absurd to many of us. Experience sees an asymmetry where fundamental physics sees none.

David Hume observed[2] that causal relations could be symmetric, what we call causes could just be events that we constantly see before what we call effects. Huw Price (1996) makes an analogous point: his conventionalist approach ("a version less arbitrary then Hume's"[3]) does not recognize any intrinsic temporal direction, and refuses to see causation as providing one. In rejecting the so-called Humean view that the asymmetry of the causal relation

1 Zeno's paradoxes where based on this: the present solution to the famous "Achilles and the Tortoise" argument (thanks to the concept of limit), after all, is not a real solution; Zeno could always insist that this is just a stipulation.

2 I am aware that there are many readings of Hume's philosophy. I am here accepting Price's interpretation.

3 Price (1996), p.137.

is merely a conventional image of earlier-later ordering, philosophers have noted that outgoing processes from a common center tend to be correlated with one another, whereas incoming processes meeting at a common center always seem to be uncorrelated (think of a video of a stone thrown in a pond, played normally and then backwards). That's why our experience strongly suggests that there is more behind this than just a temporal ordering convention – a fork asymmetry (four effects of one common cause). But the fundamental laws of physics seem to tell a different story. Consider Price (1996):

> to the extent to which there is a statistical asymmetry of this kind in the world, it is a macroscopic affair, depending on the coordinated behavior of huge numbers of microscopic components. But […] the component processes themselves seem to be symmetric in the relevant respects; there seems to be no fork asymmetry in microphysics.[4]

This fork asymmetry that we experience in the world seems to disappear when we focus on the microstructure of the physical processes in which it shows up, just as the pictorial content of an image disappears when we focus on the individual pixels that constitute it.

Imagine[5] a young man smoking a cigarette. If we see this process in the usual temporal direction, clouds of smoke come out of his mouth and disperse in the air. The reverse of this macroscopic process is, simply, absurd; but if we had a powerful zoom and saw the process at a molecular level, the situation would be radically different. A video of some molecules moving would make sense played forwards *and* backwards. It seems, the reasoning goes, that asymmetry emerges only at a macroscopic level. After all, Price argues, if we think of the history of our universe in reverse, what we would see is a collapsing universe and entropy decreasing: "there is no objective sense in which this reverse way of viewing the universe is any less valid than the usual way of viewing it. Nothing in physics tells us that there is a wrong or a right way to choose the orientation of the temporal coordinates"[6]. What Price has in mind is that a certain sequence of events (man lighting a cigarette, smoking it, tossing the butt) is happening in a certain temporal direction (past to future), which is not the only possible one.

4 Price (1996), pp. 140–141.
5 See Braibant, Giacomelli & Spurio (2009), p. 137.
6 Price (1996), p. 84.

2. Entropy and time

Within the philosophical community there is the widespread idea that the direction of time is nothing but the direction in which entropy increases. It seems, in fact, that the Second Law of Thermodynamics is the only law in physics that shows a temporal orientation. Thus, it is often considered the only candidate to physically ground the temporal orientation of our experience. Since physics, as Tim Maudlin puts it, does not distinguish at the level of fundamental laws the future direction from the past direction, "any such distinction must be grounded in contingent facts about how matter is distributed through spacetime […]. The direction of time, we are told, is nothing but the direction in which entropy increases"[7].

As I see it, there are three different possibilities here. We could claim that time is going forward *because* entropy is increasing, that entropy is increasing *because* time is going forward, or that time is going forward *and* entropy is increasing (without any causal or necessary relation between the two). The first scenario is the most accepted, because there is a physical law grounding the direction of time. The second option is preferred by philosophers with a more metaphysical taste. Their task, in turn, is to explain what grounds the thermodynamic arrow (causation, a substantivalist conception of time, etc.). The third possibility is a sort of conventionalist idea. Consider these words by Ludwig Boltzmann[8]:

> for the universe, the two directions of time are indistinguishable, just as in space there is no up or down. However, just as at a particular place on the earth's surface we call 'down' the direction toward the center of the earth, so will a living being in a particular time interval of such a single world distinguish the direction of time toward the less probable state from the opposite direction (the former toward the past, the latter toward the future).[9]

Since we constantly observe entropy increasing, we say that more events correspond to more disorder, and we call this increase of disorder *future*. This, however, is just a result of the particular point in space or time that we occupy. The Second Law of Thermodynamics, then, is not the source of the

7 Maudlin (2007), p. 17.

8 The words by the Austrian physicist serve just as an introduction to this third position: I am not maintaining that this was his view (and it probably wasn't, even if he tended to change his mind).

9 Boltzmann (1964), pp. 446–447.

temporal asymmetry, simply because there isn't any direction, any asymmetry. There is no time in a strong, Newtonian, substantivalist sense, there are only things happening in temporal relations. We are, by chance, in a particular situation in which we observe the universal entropy constantly increasing, and we call it going towards the future, but what we really mean is going towards an entropy increase. Even this Relationist version, however, share with the former two options a deep flaw: it is absolutely not clear what "entropy of the universe" means.

As popularly understood, the Second Law implies that a physical property, entropy, increases monotonically with time. Entropy, however, is definable only for systems in thermodynamic equilibrium, and the universe is *not* in thermodynamic equilibrium. Many philosophers and scientists[10], then, argue that the notion "entropy of the universe" is simply nonsense, and that the thermodynamic concept of entropy can only be defined for particular physical systems under special conditions. Thermodynamics makes sense when we focus on small thermally isolated bodies, whose volume and shape could be altered adiabatically by outside intervention. The concept of entropy itself could be understood only for systems in equilibrium, a state that cannot be ascribed to the universe as a whole. Thus, it is not clear how it is possible to ground the temporal asymmetry we see in the world on the inappropriate global extension of a local concept. This is sufficient, claims Roberto Torretti, "to dismiss the popular understanding of the second law of thermodynamics as a law of cosmic evolution, [and] to disqualify thermodynamic entropy as the physical source of universal time order"[11]; Clausius' cosmic version of the law, therefore, not only lacks any type of empirical warrant, but "clearly demands a much greater exertion of the human fancy than it is reasonable to allow in science"[12].

Obviously, we can concede that, even if it is not physically correct, it is at least intuitively possible to think of the universe as a giant box containing particles – but it is just a rough idea. The forces that are relevant at a cosmological level are completely different from the forces that are relevant for molecules in a box. Even Peter Evans, Sean Gryb and Karim Thebault, who consider – with Price – the idea of grounding our temporal experience on the entropy gradient very appealing, admit that

10 See for example Callender (2001) for a comprehensive review of the positions.
11 Torretti (2006), p. 740.
12 Torretti (2006), p. 753.

> the precise nature of the connection between the thermodynamic asymmetry and our asymmetric epistemic relationship to the past and future is a matter that remains largely unexamined […]; it is particularly problematic for the case of quantum cosmology where an explanation in terms of local thermodynamic arrow of time is inadequate. The fact that the universe is clearly not in thermodynamic equilibrium means that it is not possible to employ a thermodynamic notion of entropy in the cosmological setting.[13]

Moreover, even conceding, for the argument's sake, that the notion of universal entropy had a meaning, how should the increase of entropy determine effects to happen after causes? Imagine a box divided in two identical halves by a partition. In the left part there are ten red molecules, in the right part there are ten blue molecules, and the box is in thermodynamic equilibrium. If we remove the partition, the entropy of the system will increase (entropy could be also understood as a measure of molecular disorder within a macroscopic system). After an hour, we expect to find a confused situation. Still, it is a statistical truth that, before or after, there will be a moment at which all the red molecules will be in the left part of the box, and all the blue ones in the right. At that exact moment, with the molecular disorder to the minimum again, the experimenter will put the partition back in position. Does it mean that, in that isolated system, time passed backwards when entropy decreased? Or is it the opposite, that with the passage of time – which, in a Relationist account, corresponds to the happening of events – entropy tends to increase?

Think of a video showing what happened inside the box. It wouldn't be possible, for us, to tell whether the video is running forwards or backwards (because the beginning and the end of the video would be the same). This is true, but what's the point? It would be very inconvenient to say that when entropy increased time was running forwards, and when entropy decreased time was running backwards. Even if we observed puddles spontaneously freezing into ice cubes, why should we think that time passed backwards? There's a classical physical explanation, forwards-oriented, of what happened. Our temporal experience is intact. Simply, we have seen something very uncommon, as the particles in the box casually and spontaneously organizing into a left part and a right part, for a mere stochastic reason. We would still preserve the temporal order of the cause-

13 Evans, Gryb & Thebault (2016), p. 22.

effect relation (with causes preceding effects), and that's why I find Price's question ("could – and does – the future affect the past?"[14]) nonsensical. There's no intrinsic direction along a line in movement (forward or backward), but simply motion, events happening. Is the fact that they *could* happen (or even in fact happen) in a different order really important, or is it the very process of acquisition of information that defines, for us, a past and a future?

What I want to claim, then, is that there isn't a direction towards which time is going (there is no arrow of time), and entropy almost always increases because it is statistically simpler. Eternalist accounts that make use of an arrow of time usually envisage a Block universe in which, once we have chosen a point *and* a direction, we can tell what is going to happen: they take an outside perspective and see the entirety of spacetime as an immobile cinematographic film, where all the spatiotemporal stages are spread out, and we have to choose in which direction we want to play the film. What I'm proposing, instead, is to think that, whichever freeze-frame we choose, only two things are possible, to move or not to move. We can consider, from the external perspective, the freeze-frames, but whichever moment we pick, the world has no choice but moving, consuming energy, getting older.

3.The low-entropy past and the banality of chaos

If the entropy of the universe is still increasing, and entropy is a measure of the disorder of a system, how come that the beginning of our universe was so orderly? There are two main problems with this question: the first is that, as I have argued, the expression entropy of the universe, simply, is meaningless – the universe is not a box of particles. The second is that we still don't have – and maybe we will never have – a clear understanding of what "the beginning of the universe" means. A part from the old scholastic question "why is there something instead of nothing?", even the contemporary, qualitative description of the beginning of the universe involve infinite quantities (singularities), which are maybe good ingredients for a mathematical description of reality, but definitely not for a physical one. What does it mean that all the matter and the energy of the universe were concentrated in one point? Does it really make sense to add to this story the fact that matter and energy had also a very low entropy? All we know (or at least believe) is that after the Big Bang there was a rapid expansion and clumps of matter were formed, which are still exchanging energy with the surrounding environment.

14 Price (1996), p. VII.

This is related to the question of the alleged fine tuning of the fundamental parameters of the universe. I suspect, in fact, that a similar reasoning is responsible for the emergence of the problem of the Past Hypothesis in thermodynamics[15]. If we think of all the circumstances that led our parents to meet, fall in love and have a child at a particular moment, we realize that the probability for us to be born was one in billions. It seems that everything in the universe secretly conspired to make our birth happen - every random action seems extraordinarily *ad hoc* for the organism at the end of the causal chain. The mere fact that our parents could have had a different child doesn't mean that we have to explain why they had us and not our possible brothers. Saying that it was random isn't concealing a deeper truth.

Here, in my opinion, we are underestimating the banality of chaos. The Second Law of Thermodynamics, after all, tells an obvious story. I'm not sure it should even be considered a law of nature, it is merely a statistical truth. Think again of the clouds produced by the man smoking. If you concentrate on the single particles, you lose track of the direction of time, while the same thing is impossible when you consider the entire picture. This is true, but it has a simpler explanation than that of an arrow of time pointing in a certain direction. Consider a lottery (90 possible numbers, 1–90, and six extractions) and these two different extractions:

13 41 2 87 60 35
1 2 3 4 5 6

Obviously, the two sequences have the same probability to be extracted. But there is a clear sense in which the second sequence strikes us as incredible (that's probably why, even if it doesn't make sense, no gambler in the world would ever spend a penny on the second sequence). When we concentrate on the single numbers, we lose track of this amazement. The six (final number drawn), in itself, isn't a shocking result. When we look at the big picture, however, the six strikes us as the perfect fulfillment of a miracle. Why is that so? Because an ordinate and dense sequence is extraordinarily much more improbable that a random one. There are a lot of disordered sequences, that is to say, while there are only a few ordinate ones. The first sequence is just as rare and incredible as the second one, but there is a sense in which it is different, it is part of a much-populated class of random sequences. Even if the two sequences have the same probability to be extracted, that is to say, it

15 See Callender (2001).

is extraordinarily much more probable to obtain a random sequence than an ordinate one, simply because the random sequences are a lot more. I suspect that this is exactly what happens in the case of the man smoking.

If we concentrate on the microphysical events, the single particles moving, we lose track of the big picture, just as if we only considered a number at a time in the case of the lottery. In this sense, it is true that from a physical point of view the process is time-reversal invariant, just as in the case of the lottery it is indifferent to extract a six or a seventy-one. But when we consider the macrophysical situation, it is much more probable that the particles disperse in the air. There is a clear sense in which there is no need of a particular law of nature to see the particles disperse instead of gather, it is simply a matter of statistics. Chaos is infinitely much simpler than order, and the Second Law of thermodynamics simply states this statistical truth, it doesn't impose a direction on nature. That's why we almost never observe the decrease of entropy in an isolated system.

The whole entropic arrow argument seems based on a confusion between the physical possibility to see a very rare process (the opposite of a physical process we are accustomed to), and the alleged possibility to see one and the same process in two different temporal directions. The so-called time reversal invariance of the fundamental laws of physics should be read just as the possibility, for the twenty molecules in the box described above, to casually order themselves spontaneously. It has nothing to do with a direction of time.

This, in turn, is related to the debate about causality. We always see the flame *after* we rub the match. The actual laws of physics seem to inherit Hume's worry that, however, there is nothing forbidding the opposite, that the flame could be the cause of the rubbing of the match. It seems that everything relies on initial conditions, which, the reasoning goes, are not part of the laws of nature, but just define the physical state to which they apply. In a different universe with different initial conditions, its inhabitants could be used to the opposite of what we are used to see, and maintain that it is absurd to think that the rubbing precedes the flame.

It seems that there is no inbuilt asymmetry in the laws of physics, and physical laws only yield concrete predictions when they are coupled to particular boundary conditions, typically formulated as initial conditions. If we assume very special initial conditions, the time-symmetrical laws might still allow only a time-asymmetrical solution. Following this strategy, the direction of the cause-effect relation would not be law-like but due to a contingent feature of our universe. But, as observes Maudlin (2007),

> the laws of nature *alone* suffice to explain almost nothing […]. The models of fundamental physical law are infinitely varied, and the only facts that those laws *alone* could account for are facts shared in common by all the models. In all practical cases, we explain things physically not merely by invoking the laws, but also by invoking *boundary conditions*.[16]

The fact that the reversed order of a physical process is possible doesn't mean that the actual order in which it happens isn't objective. It does not even mean however, as Maudlin implies, that there is an objective, Substantivalist-like, "intrinsic passage of time"[17]. This is a crucial point:

> the motion of an asteroid from Earth to Mars is just a matter of the asteroid being differently situated with respect to those planets at different times […]. Since these are objective, mind-independent facts about space-time worms, the changes are equally objective and mind-independent. The rub, of course, is that the asteroid being differently situated at different times is consistent both with a motion from Earth to Mars and with a motion from Mars to Earth […]. Motions and changes are not merely a matter of things being different at different times, but also critically a matter of which of these times are *earlier* and which *later*. […] If there is no difference in the entropy (e.g. if the universe is in thermal equilibrium), then there is no longer a distinction between Earth-to-Mars and Mars-to-Earth trips.[18]

The travel of the asteroid as we experience it, however, is related to changes in my body (for example, me getting older) that clearly define a trajectory, without the need of a Substantivalist notion of time or the existence of an arrow. The asteroid being differently situated at different times is consistent both with a motion from A to B and the opposite, but a relational notion of change is able to distinguish between the two trips. If I saw the asteroid near Earth when my hair was brown (young man) and near Mars when my hair was white (old man), this clearly defines an earlier and a later. This objective, mind-independent order defines an objective relational time, different to Newton's in that it depends on the actual change of the things that surrounds

16 Maudlin (2007), p. 119.
17 Maudlin (2007), pp. 127–128.
18 Maudlin (2007), p. 128.

us, and doesn't pass *in se et per se*. There is only motion, there isn't a Substantivalist time passing, nor an arrow of time. All we need in such a Relationist account are things moving.

4. There is no arrow of time

As I have argued, the time-reversal invariance of the fundamental laws of physics is often taken to mean that in a Block Universe, even if we are experiencing events as forwards-in-time, it would be perfectly possible to experience them symmetrically, as backwards-in-time. But what does it really mean?

Even setting aside the observed violations of charge parity invariance in the decay of the neutral K meson (a counterexample to the alleged time reversal invariance of the fundamental laws of physics), and supposing that we can understand the time reversal operation without there being an objective direction with respect to which the reversal occurs (you can play a video forwards or backwards, but what about the events you filmed? They must have happened one after the other), what I find really troubling is the idea that events could happen in the opposite temporal direction.

Consider the spatial case. You can move or not move, you can't undo your movement. If you walk 80 meters, you can certainly come back. The result, however, is not 0, but 160 meters. A video, in this case, would just reproduce the illusion I described for the temporal dimension. Playing it backwards, it seems that the person who walked 80 meters could go back to 0. But this is just wrong, moving is always positive.

I believe that this is also the case with the Relational notion of time that I'm trying to defend. Either time passes or not, it can't go in another direction. If we intend time as a measure of change, the world has only two possibilities, stay still or move – just like us when we walk. What I want to claim, then, is that even temporally all we can do is to add meters to our walk. We can not subtract, because there is not a direction towards which time is flowing. As the Shakespearean Hamlet would put this: to pass or not to pass (which, in a relational sense of time, means for things to change, to move). The only difference between time and space, under this point of view, seems to be the fact that while I can stay still in space, I can't do that in time. But this is true only at a common-sensical level. Even when we sit down, we are moving really fast in space (around the terrestrial axis *and* the sun). In General Relativity, moreover, the clear distinction between time and space seems to vanish, moving is always moving in time *and* space. But even if, for

the sake of the argument, we kept the conceptual distinction between space and time, it would be possible, although very difficult, to stay still in time (we should totally freeze the universe, as in the famous thought experiment by Sidney Shoemaker, 1969).

If we re-think the whole question under this light, the time reversal invariance of the fundamental laws of physics could be read just as the theoretical possibility that the order of a certain chain of events was different, and not as the possibility that the same order of events happened in a different temporal direction, because there is no such thing as a positive (or negative) temporal direction. As Maudlin points out, the actual possibility of the reversed order of the physical processes we usually see is not to take for granted. It would mean that

> given the actual sequence of physical states of your body over the last ten minutes, the time-reversed sequence of time-reversed states is also physically possible. Somewhere on some other planet (as far as the laws of physics go) some such sequence could exist, unproblematically time reversed relative to the sequence of states that make you up [we can label this strange order of physical states the Doppelgänger point of view]. The visual system* of the Doppelgänger is […] quite unusual: rather than absorbing light from the environment, the retina*s emit light out into the environment. (The emitted light is correlated with the environment in a way that would seem miraculous if we did not know how the physical state of the Doppelgänger was fixed: by time-reversing a normal person.) […] There is no reason to belabor the point: in every detail, the physical processes going on in the Doppelgänger are completely unlike any physical processes we have ever encountered or studied in a laboratory, quite unlike any biological processes we have ever met.[19]

Either the Doppelgänger has a mental state identical to ours, but then it hasn't a different perspective, or the physical processes going on are completely unlike biological processes we have ever met (more magic than science). I take the moral to be that it is an error to concentrate on the microphysical world, forgetting the big picture. As I argued, we would fail to recognize the incredible improbability of an ordinate sequence in the lottery, which is

19 Maudlin (2007), p. 123.

different from the chance of the single sequence – which never changes. Focusing on the particular is not a neutral operation.

Even if we think that we are in a Block Universe and past, present and future events are real, that doesn't mean that we are going from left to right along a temporal line, or that it would be possible to go from right to left. My point, then, is that there isn't an arrow of time, there is not a direction in which time is passing, and thus not any alternative direction. Events are happening – even if I see the disordered particles in the box reorganizing in their respective half, I have seen *something more*, I am older. Why should I think that time is passing backwards? Simply, in an unlikely case like that, entropy would demonstrate its stochastic nature, it would just be another sign of the fact that entropy and time are not related.

To understand why the argument I am trying to make is not circular, please consider the following passage by Huw Price (1996):

> as Boltzmann himself saw […] there is *no* asymmetry […]. The above point about entropy increase toward (what we call) the future applies equally toward (what we call) the past. At a given starting point there are very many more possible histories for the gas that correspond to higher entropy macrostates in its past, than histories that correspond to lower entropy macrostates. Insofar as the argument gives us reason to expect entropy to be higher in the future, it also gives us reason to expect entropy to have been higher in the past. Suppose we find our gas sample unevenly distributed between its two chambers at a particular time, for example. If we consider the gas's possible future, there are many more microstates which correspond to a more even distribution than to a less even distribution. Exactly the same is true if we consider the gas's possible past, however, for the statistical argument simply relies on counting possible combinations, and doesn't know anything about the direction of time.[20]

This is exactly the point. If the past is considered as a temporal locus from which we are moving away, we would be forced to admit that we are actually going in a direction. But what if we, much more radically, considered past every moment at which there were fewer movements (at which our body was less ruined)? It is the charge of circularity itself, at this point, that becomes circular. I am claiming the most parsimonious thing, which is that fewer

20 Price (1996), p. 30.

movements (a less ruined, used brained) correspond to the past. There is no line, no direction, only motion. It would be Price, in this case, that would have to explain in which sense a baby could be in the future of an old man, it would be Price that would have to presuppose the existence of an arrow of time. Why should we describe time as flowing along a uni-dimensional line? It is way simpler a model in which 'past' and 'future' stand for 'fewer movements' and 'more movements'. Let me explain that with a thought experiment.

Imagine an extra-temporal, omnipotent god able to completely stop the change and the motion in the universe and, if you believe that time is a flowing independent entity, stop time. Suppose that this god concentrates on a hot bar of iron on earth, which was placed in a cold box of metal just before the complete stop. If the god did not change anything, the bar would not distribute its heat during this complete stop (since, to do so, the atoms would need to move); but what if the god decided to instantaneously and casually mix the atoms inside the hot-bar-and-cold-box system? For example, suppose he decided to throw three dice – which he created in such a way that he can not predict the result. The first time, the three dice individuate an atom in the system. The second time, the dice tell the god which atom to exchange for the first one. After the extra-temporal god has done this for several times, the entropy of the system is very probably increased, but the world hasn't gone in any direction, it hasn't gone anywhere, indeed. The god, per hypothesis, is extra-temporal, his actions are not going from past-to-future or from future-to-past, they are instantaneous. It doesn't seem that the increase of entropy is connected with a particular direction in which we choose to play the film of reality, but simply to a casual change. It's merely a stochastic reason, it is simpler to disorder a deck of card than to order it. We don't need an arrow of time to explain that.

If we think that there is an arrow of time, we have the problem of explaining why effects always precede causes, why the arrow points in a certain direction. But if we take seriously my proposal, the problem doesn't exist. It would just be a matter of fewer/more movements, and we don't need a direction for that. An old man has done more movements than a baby, it is not a mystery that his consciousness is aligned in the direction baby-to-old man. Do we need an arrow of time for that?

From an Eternalist point of view, there is a four-dimensional, unchanging world. But when you consider a particular point of view (a particular spatio-temporal point), either there are zero degrees Kelvin, or it moves, it changes,

it wears out. An atom isn't moving because time is passing, it moves because it has a mass and an energy. It doesn't go in a temporal direction, it simply moves! Does the fact that, from a microphysical point of view, its reversed movements are also possible, automatically imply that it is going towards a specific temporal direction?

What we call time has always a fundamental reference to motion. A day is a rotation of Earth on its axis, a second is a certain number of oscillations of a Cesium atom, and so on. Our consciousness moves because the atoms in our brain move, they don't go in a direction, and there is always a clear sense in which a 70-years-old brain is older than a 20-years-old brain (it is more ruined) that doesn't make reference to the passage of time in itself, or its direction, or its arrow.

The reason why we feel like our consciousness is moving forward is that at every point our consciousness change, with the last acquisition of data. If we think of that as a movement along a line, we naturally think that we are going in a direction. But as I should have shown, there is also the possibility to think that the only two options are to move or not to move. There is no movement of our consciousness on a temporal line. There is just the motion of atoms and the rate at which our brains capture changes in respect to that motion.

Conclusions

In the first paragraph, I claimed that the mere theoretical possibility of time reversal invariance of the fundamental laws of physics is something strongly related to a mathematical, misleading description of our universe. In the second, I maintained that entropy and time are not related, and that the notion of entropy of the universe has many problems in itself. In the third, I argued in favor of the banality of chaos. The Second Law of Thermodynamics, far from being a law of nature, simply states a stochastic truth. In the fourth, I sketched an account in which change is not going in a direction, and defended it from some possible replies.

Some Eternalists claim there is a Block, and we – or our consciousness – are traveling along it in a particular direction, but it is physically possible to also travel in the opposite direction. I answered that there isn't any direction, any arrow. It is simply a mathematical fiction, resulting from the focus on the microphysical particulars instead of the big picture, failing thus to see the banality of chaos. It seems to me the most natural thing is to claim that events simply happen. The fact that they could have happened in a different order

does not entail that they happened in a certain direction instead of another. Whichever atom in spacetime you choose, if it has an energy it moves, it is not going in a temporal direction. Many atoms moving randomly result, for mere stochastic reasons, in macro-situations of increasing chaos. There is no direction towards which they are going, there is just moving or not moving, and from a Relationist point of view, motion is not the result of a mysterious independent passage of time, but the passage of time itself.

References

Arntzenius, F. (2012). *Space, Time, and Stuff*. Oxford: Oxford University Press.

Boltzmann, L. (1964). *Lectures on gas theory*. Berkeley CA: University of California Press.

Braibant, S., Giacomelli, G., Spurio, M. (2009). *Particelle e interazioni fondamentali: il mondo delle particelle*. Springer.

Callender, C. (2001). Thermodynamic asymmetry in time. *Stanford encyclopedia of philosophy*.

Dainton, B. (2010). *Time and Space (second edition)*. Durham: Acumen Publishing Limited.

Evans, P., Gryb, S., Thebault, K. (2016). ψ-epistemic quantum cosmology? *Studies in the History and Philosophy of Modern Physics*, Vol. 56, pp. 1–12.

Maudlin, T. (2007). *The metaphysics within physics*. Oxford: Oxford University Press.

Morganti, M. (2017). Relationism about time and temporal vacua. *Philosophy*, Vol. 92, pp. 77–95.

Price, H. (1996). *Time's arrow and Archimedes' point*. Oxford: Oxford University Press.

Shoemaker, S. (1969). Time Without Change. *Journal of Philosophy*, Vol.66, pp. 363–381.

Skow, B. (2015). *Objective Becoming*. Oxford: Oxford University Press.

Torretti, R. (2006). The problem of time's arrow historico-critically reexamined. *Studies in History and Philosophy of Modern Physics*, Vol. 38, pp. 732–756.

Andrea Roselli
roselli.uniroma3@gmail.com
Università degli Studi di Roma Tre
Dipartimento di Filosofia, Comunicazione e Spettacolo
Via Ostiense 234, Roma (00154) Italy

Defusing the Miners Paradox

Michael Shaffer

Abstract
This paper presents a case for the claim that the infamous miners paradox is not a paradox. This contention is based on some important observations about the nature of ignorance with respect to both disjunctions and conditional obligations and their modal features. The gist of the argument is that given the uncertainty about the location of the miners in the story and the nature of obligations, the apparent obligation to block either mine shaft is cancelled.

1. Introduction

In this paper a more nuanced and accurate construal of the miners paradox is presented and on this basis the Miners paradox is defused. This involves understanding some important points about rational obligation, disjunction and uncertainty. The main contentions made here are based on the observation that crucial modal and epistemic dimensions of the story are totally absent in typical presentations of the paradox. Specifically, these modal and epistemic dimensions are left out of the typical formalizations of the disjunctive knowledge involved and the conditional obligations that are at the heart of the alleged paradox. When these notions are included in the formal translation of the story and when they are added in it turns out that there is no paradox in the miners' story at all. This manner of dissolving the miners paradox is to be preferred to alternative solutions—particularly that of Kolodny and MacFarlane (2010)—on the basis of its relative simplicity. Importantly, it does not require radical revisions of the logical of indicative conditionals and the rejection of the unrestricted validity of modus ponens, as Koldny and MacFarlane's solution requires. Let us begin then by focusing on some important aspects of knowledge as they pertain to disjunction.

Filosofiska Notiser, Årgång 5, Nr. 1, Maj 2018, 57–67

2. Disjunctions and Uncertainty

Consider the following story:

> Joe wakes up the day after the 2000 U.S. presidential election. He has not followed the details of the race and knows only that there were two candidates being voted for: George W. Bush and Al Gore. He is aware that only one of them could have won, he does not know which one won and has no evidence to favor either the claim that Bush won or the claim that Gore won. Joe meets up with his buddy Tony and Tony asks Joe "Who won the election?" Joe responds with "Either Bush or Gore."

In light of this brief story, consider the following parsing of Joe's assertion, where we understand clearly that the component sentences involved are contingent:

> (BG) Either Bush won the 2000 U.S. presidential election or Gore won the 2000 U.S. presidential election.

This ordinary language English sentence might be understood to have the following richer correlate:

> (BGA) Either Bush *actually* won the 2000 U.S. presidential election or Gore *actually* won the 2000 U.S. presidential election.

More formally, where W stands for "Bush won the 2000 U.S. presidential election" and G stands for "Gore won the 2000 U.S. presidential election," we can regiment BGA simply as follows, where "■" is an actuality operator and "∨" is standard disjunction:

> (BGA1) ■W ∨ ■G

Additionally, and as explicitly noted in the Bush/Gore story, it is clear that the sort of disjunction involved does not permit it to be that case that Bush won and Gore won, so we need to amend things as follows based on the recognition that ¬◊(W & G), where "&" is standard conjunction and "¬" is standard negation:

(BGA2) $(\blacksquare W \vee \blacksquare G) \& \neg(\blacksquare W \& \blacksquare G)$.

In ordinary discourse the use of sentences like BGA also seems to connote epistemic uncertainty with respect to the truth of the disjuncts involved.[1] If this is the case, then with respect to Joe, BGA can be regimented as follows, where $K_J x$ is "Joe knows that x":

(BGA3) $(\blacksquare W \vee \blacksquare G) \& \neg(\blacksquare W \& \blacksquare G) \& (\neg K_J \blacksquare W \& \neg K_J \blacksquare G)$.

This is just the claim that one of either Bush or Gore actually won the election but Joe does not know which of Bush or Gore actually won the election. Finally, there is an implicative connotation involved that it is possible that Bush won and that it is possible that Gore won and that the utterer knows this. The disjunction concerning the actuality of Bush winning or the actuality of Gore winning is not supposed to be true in virtue of the fact that one of the disjuncts is impossible and the other actually true. So the epistemically complex, modalized, use of disjunction in such contexts suggests the following rather complex rendering, where "◊" is the orthodox possibility operator of modal logic:

(BGA4) $(\Diamond W \& \Diamond G) \& (K_J \Diamond W \& K_J \Diamond G) \& (\blacksquare W \vee \blacksquare G) \& \neg(\blacksquare W \& \blacksquare G) \& (\neg K_J \blacksquare W \& \neg K_J \blacksquare G)$.[2]

This appears to be a typical rending of the epistemically and modally rich use of "or" in cases involving contingent statements where there is uncertainty and this has important implications for the Miners paradox, which has received much attention of late in the context of both ethics and epistemology.[3] What it suggests is that the ordinary language usage of disjunction involves important epistemic and modal content that is overlooked in the standard logical translations of disjunction.

[1] This is clearly the case when the epistemic agent in question does not know the disjunction to be true on the basis of knowing that one of the disjuncts is true but not knowing anything about the truth value of the other or does not know the disjunction to be true on the basis of knowing one disjunct to be true and employing weakening by disjunction introduction. The view of disjunction developed here is then closely related to the modal account of disjunction developed by Zimmerman (2000) and Geurts (2005) and inspired by Kamp (1973).

[2] Again, this view is then closely related to that defended in Zimmerman (2000) and in Geurts (2005).

[3] See Parfit (manuscript), Kolodny and MacFarlane (2010) and Dutant and Fitelson (manuscript).

3. The Miners Paradox

Here is a version of the story that gives rise to the miners paradox. It is essentially the same as that presented, for example, by Kolodny and MacFarlane (2010):

> Ten miners are trapped in a flooding mine; they are either all in shaft A or all in shaft B. Given Tony's information, each location is equally likely. Tony has just enough sandbags to block one shaft. If the miners are in the blocked shaft, they will all be saved. If the miners are in the other shaft, then they will all be killed. If Tony does nothing, the water will distribute between the two shafts, killing only the one miner at the lowest level of the mine.[4]

The paradox implicit in this sort of story is supposed to be derived from considerations raised initially by Jackson (1991) about possible cases where the alternative with the best outcome does not have the highest utility on the body of known information in the situation so described. Let us then turn to the presentation of the alleged paradox.

On the basis of this story the following claims seem to be true:

(M1) Tony ought to block neither shaft.
(M2) If the miners are in A, Tony ought to block A
(M3) If the miners are in B, Tony ought to block B.
(M4) Either the miners are in A or they are in B.

(M2)–(M4) entail,

(M5) Either Tony ought to block A or Tony ought to block B.

Prima facie, the paradox and the disjunctive uncertainty involved in the miners story can be formally regimented as follows, where $O_T x$ is "Tony is rationally obligated to do x",[5] M_A is "miners are in A", M_B is "miners are in B", B_A is "block shaft A", B_B is "Block shaft B and "$\rightarrow$" is standard implication:

[4] See also Parfit (manuscript), Regan (1980) and Pettersson (2014).
[5] "Rational obligation" is just meant here to indicate some rationally mandated action that follows from one or more rational principles.

(M*1) $O_T\neg B_A$ & $O_T\neg B_B$.
(M*2) $M_A \rightarrow O_TB_A$.
(M*3) $M_B \rightarrow O_TB_B$.
(M*4) $M_A \vee M_B$.
(M*5) $O_TB_A \vee O_TB_B$.
(M*6) $\neg\Diamond(B_A$ & $B_B)$.
(M*7) $\neg(\blacksquare M_A$ & $\blacksquare M_B)$.
(M*8) $\Diamond B_A$ & $\Diamond B_B$.[6]

So, there appear to be conflicting obligations in this case. Technically, this is not a contradiction. Generating a contradiction from M*1–M*5 requires additional steps as follows:

(M*9) $O_T\neg B_A \rightarrow \neg O_TB_A$.
(M*10) $O_T\neg B_B \rightarrow \neg O_TB_B$.

M*1, M*9 and M*10 imply this:

(M*11) $\neg O_TB_A$ & $\neg O_TB_B$.

Given these additional steps the miners paradox then at least appears to be a bona fide paradox.

4. Solution

However, even this appearance is deceptive in light of the epistemic modalities and uncertainties involved in the use of "or" in the story and in the parsing of the conditional obligations that are crucial to the story. This comports with the point made above about the ordinary language implicature associated with disjunction. When we incorporate these facts into the regimentation of the paradox we get the following, more complex characterization of the miners paradox propositions:

(M*1) $O_T\neg B_A$ & $O_T\neg B_B$.
(M*2′) $(\blacksquare M_A$ & $K_TM_A) \rightarrow O_TB_A$.
(M*3′) $(\blacksquare M_B$ & $K_TM_B) \rightarrow O_TB_B$.

[6] M*6, M*7 and M*8 are not required for deriving the miners paradox, but, given the account of disjunctive uncertainty proposed here, they are parts of the story. Moreover, they are crucial parts of the solution to the alleged paradox of the miners.

(M*4′) $(\blacksquare M_A \vee \blacksquare M_B) \,\&\, (\neg K_T M_A \,\&\, \neg K_T M_B) \,\&\, (K_T \Diamond M_A \,\&\, K_T \Diamond M_B)$.
(M*5) $O_T B_A \vee O_T B_B$.
(M*6) $\neg \Diamond (B_A \,\&\, B_B)$.
(M*7) $\neg (\blacksquare M_A \,\&\, \blacksquare M_B)$.
(M*8) $\Diamond B_A \,\&\, \Diamond B_B$.[7]

Notice here that the important changes are to be found in M*2′, M*3′ and M*4′ and they importantly involve discrimination of what is actually the case and what is known to be the case in the story. The conjunction of M*4′, M*7 and M*8 is analogous to BGA4 and it reflects the same kind of context involving the use of disjunction in light of epistemic modalities and uncertainty we found in the Bush/Gore story. As such, it is appropriate to make these changes in the same way. As per the miners' story then, the agent involved in the situation does not have information favoring either $\blacksquare M_A$ or $\blacksquare M_B$ and so does not know M_A and does not know M_B, although he knows both M_A and M_B are possible and that just one must be actually true.[8] Thus, M*4′ captures better the position of that agent with respect to the disjunction involved in the miners story and his/her knowledge with respect to the disjuncts.

Crucially then, the next contention made here is that in M*2′ and M*3′ the rational obligation to block one but not both of the mine shafts is conditionally dependent on the agent's *knowing that M_A or M_B*, respectively.[9] In other words they are subjective obligations. This bit of absolutely fundamental information is absent in the initial presentation of the putative paradox. It is, however, an entirely plausible and principled assumption. If the miners are in shaft A or in shaft B respectively, but Tony does not know this, then Tony cannot reasonably be thought to have an obligation to block that shaft in question rather than the other. To deny this principle in general would impose a plethora of unknown and/or unknowable obligations on every agent, the resultant objective obligations would be practically worthless in deliberations about what to do under conditions of limited information and it would involve violations of a plausible version of the "ought implies can" principle. This latter point follows because such obligations would be obligations that the agent *could not*—in the sense of epistemic

[7] The putative paradox can also be presented in terms of justified belief rather than knowledge, but this changes nothing about the analysis of the case and about the solution proffered to it here.
[8] Moreover, he has not reasons to favor the truth of M_A over M_B or M_B over M_A.
[9] See Fischer and Ravizza (1998), Ginet (2000), Rosen (2008) and Mele (2011) for philosophical defenses of this claim.

possibility—meet in light of this sort of ignorance. They would be obligations that are epistemically impossible to meet in the sense that they are required of the agent even though the agent is totally unaware of them. Simply consider the following scenario involving such a conditional obligation:

> Bill is taking a walk in the woods. He is near a river, but his view of the river is totally obscured. It is also loud. So, he cannot see or hear anything in the river. Sally, who cannot swim, has fallen in the river near enough to Joe so that he could physically reach her and easily save her. He is an excellent swimmer.

If obligations like those that are alleged to pertain to the miners in the simple version of the putative paradox are legitimate, then by parity of reasoning we would have to say that Bill is obligated to save Sally where he could easily save her, but where he has no knowledge that she is drowning. But it is clear that his epistemic state defeats the conditionality of that obligation and it is not even remotely plausible to claim that he does, in fact, have such an obligation. He is exculpated from that obligation in virtue of his ignorance (which is no fault of his) and this is an utterly typical but reasonable sort of excuse in such cases. Specifically, the conditional obligation is defeated by such ignorance and such agents rightly can claim that they had no such obligation when the agent is unaware of it and is not at fault in being unaware of it.[10] The same thing then goes for the miners paradox conditional obligations. As such, where $O_SR\neg p$ is "S should bring it about that $\neg p$" and p is a factual state with negative consequences such that it is morally bad that p, the position defended here is that obligations with the general form $(p \ \& \ \neg K_Sp) \rightarrow O_SR\neg p$ are not (at least not always) obligations and that reasonable obligations (at least typically) have the form $(p \ \& \ K_Sp) \rightarrow O_SR\neg p$.[11] But, notice then that given the epistemic uncertainty involved in M*4′ and this more accurate rendering of the obligations involved, the inference from M*2′, M*3′ and M*4′ to M*5 is invalid. One cannot derive a rational, subjective, obligation to block A or to block B from the fact that the miners *might be in A* and *might be in B* when the agent does not know either

[10] This contention also has strong empirical support as Kissinger-Knox, Aragon and Mizrahi (2018) demonstrates.

[11] See Spencer and Wells (forthcoming) for discussion and defense of rational obligations and knowledge requirements like the one suggested here.

possibility to be the case and does not have any reason(s) to favor the truth of M_A over M_B or of M_B over M_A.[12] This is the very kind of ignorance that defeats the obligations involved. So, there is no paradox here and given M*1, M*2′, M*3′ and M*4′ it is clear that, rationally, Tony ought not to block shaft A and ought not to block shaft B in light of his ignorance about the location of the miners.

The last matter that needs to be addressed here concerns Kolodny and MacFarlane's (2010, 118–119) rather convoluted contention that we cannot properly interpret the conditional obligations involved in the miners paradox as subjective obligations in the manner suggested here. They do so on the basis of the claim that subjectivist interpretations of conditional obligations "…cannot make good sense of the use of "ought in advice" (Kolodny and MacFarlane 2010, 119)." That is to say that if such obligations were subjective in the sense suggested here, we would not be able to make sense of advice about situations where advice is given by agents in superior information states to agents in inferior information states with respect to the very same situation. Consider their Dialogue 1, where the agent in the miners paradox has an exchange with an adviser who, *ex hypothesi*, knows the location of the miners:

> AGENT: I ought to leave both shafts open, guaranteeing that nine survive.
> ADVISER: No, you ought to block shaft A. Doing so will save all ten of the miners (Kolodny and MacFarlane 2012, 119).

Kolodny and MacFarlane contend that the subjective construal of conditional obligations cannot properly make sense of this sort of exchange for the following reasons.

They claim, first, that the subjectivist construal of conditional obligations makes sense of Agent's assertion in Dialogue 1 given his/her limited information. But, second, they contend that this is not true of Adviser's assertion in Dialogue 1. With respect to this scenario, they rightly claim that Adviser is not making a claim about what Agent ought to do *given Agent's limited information*, for Agent already knows that and then Agent and Adviser would not be in disagreement when Adviser challenges Agent's assertion. In order to make sense of this, Kolodny and MacFarlane claim that

[12] Notice that the conditional obligations also do not follow where the undesirable state is known merely to be possible.

the advice from Adviser only makes sense if there is real disagreement between Agent and Adviser. Adviser knows where the miners are located and so challenges Agent's claim to be obligated not to block either shaft and, according to Kolodny and MacFarlane, this makes sense only if Adviser is really in disagreement with Agent. After claiming that this desideratum cannot be met, they argue further that the subjectivist reading cannot be saved by claiming that Agent acquires evidence about the location of the miners upon hearing Adviser's claim and so is no longer obligated to not block the shafts due to Agent's ignorance of the location of the miners. But, this latter argument is really a bit of a red herring, for it is easy to make sense of the disagreement between Agent and Adviser in terms of the subjectivist construal of conditional obligations.

First, let us refer to Agent as "Tony" and Adviser as "Vivian". So, on the subjectivist reading, in Dialogue 1 Tony has the following obligations: $O_T\neg B_A$ & $O_T\neg B_B$. This is only the case, however, because of his ignorance of the location of the miners as we saw previously and regimented as follows: $(\blacksquare M_A \vee \blacksquare M_B)$ & $(\neg K_T M_A$ & $\neg K_T M_B)$ & $(K_T\Diamond M_A$ & $K_T\Diamond M_B)$. The crucial subjectivist bases for the conclusion that Tony ought to block neither shaft are the following claims: $(\blacksquare M_A$ & $K_T M_A) \rightarrow O_T B_A$ and $(\blacksquare M_B$ & $K_T M_B) \rightarrow O_T B_B$. However, Vivian's situation is entirely different. From the perspective of her information state the following claims are true: $O_V\neg B_A$, $K_V M_A$, $\blacksquare M_A$ and $(\blacksquare M_A$ & $K_V M_A) \rightarrow O_V B_A$. Vivan's obligation would then be entirely different from Tony's if she were in a position to act to save the miners and knows what she knows. But, Tony does not know the location of the miners and so does not have the same conditional obligation as Vivian. Kolodny and MacFarlane appear to reject subjectivism, at least in part, on the basis of the following utterly implausible claim: $(\blacksquare M_A$ & $K_V M_A) \rightarrow O_T B_A$. More importantly, they contend on this basis that the subjectivist *cannot* explain the sense of disagreement in Dialogue 1. But this is simply not true. The disagreement between Tony and Vivian is easy to understand in terms of subjectivism about conditional obligations, independent of any worries about how Vivian's assertion effects Tony's evidential state.

Vivian disagrees with Tony about what Tony should do, but only in the sense that *Tony's obligations* can be understood relative to two possible information states that *Tony* could be in and only one of which he is actually in. So, Vivian is simply disagreeing with Tony in the sense that she is saying something like this: "No. You should block shaft A, because *if you knew what I know* that would be your correct obligation." But, Tony's information

state is such that he does not know what Vivian knows. If he in fact learns what Vivian knows, then his correct conditional obligation would change with the alteration in his information state and it would (at least in normal cases) conform to the subjective obligation that Vivian herself has given her information state (i.e. to block shaft A). So, they disagree about what is the right thing to do only in the sense that they derive different subjective obligations for Tony, but they do so on the basis of the different information states Tony could be in. What Vivian is saying is simply that Tony could be in a better state of information (one that she, in fact, occupies) and if that were the case, then Tony would no longer be obligated to block neither shaft. Of course, he is not in that state though in Dialogue 1 and since he isn't in that information state he does not have the obligation to block only shaft A. He is exculpated from the obligation to do that because of his impoverished information state, and this would not be the case if he were in Vivian's information state. Vivian's advice then is nothing more than the specification of an epistemic possibility that is not currently an actuality for Tony and her advice is nothing more than an entreaty to Tony to improve his information state. So, this objection does not really undermine the subjectivist interpretation of conditional obligations. The sense of disagreement involved in Dialogue 1 is simple to understand and given the reasons in favor of the subjectivist interpretation of conditional obligations discussed earlier, the solution to the miners paradox presented here is to be preferred to alternatives that are far less conservative.

References

Dutant, J. and Fitelson, B. (manuscript). Knowledge Centered Epistemic Utility Theory. http://fitelson.org/boulder.pdf.

Fischer, J. and Ravizza, M. (1998). *Responsibility and Control*. Cambridge: Cambridge University Press.

Geurts, B. (2005). Entertaining Alternatives: Disjunctions as Modals. *Natural Language Semantics* 13: 383–410.

Ginet, C. (2000). The Epistemic Requirements for Moral Responsibility. *Nous* 34: 267–277.

Jackson, F. (1991). Decision-Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics* 101: 461–82.

Kamp, H. (1973). Free Choice Permission. *Proceedings of the Aristotelian Society* 74: 57–74.

Kissinger-Knox, A, Aragon, P. and Mizrahi, M. (2018). Does Non-Moral Ignorance Exculpate? Situational Awareness and Attributions of Blame and Forgiveness. *Acta Analytica* 33: 161–179.

Kolodny, N. and MacFarlane, J. (2010). Ifs and Oughts. *Journal of Philosophy* 107: 115–43.

Mele, A. (2011). Moral Responsibility for Actions: Epistemic and Freedom Conditions. *Philosophical Explorations* 13: 101–111.

Parfit, D. (manuscript). What We Together Do. http://individual.utoronto.ca/stafforini/parfit/parfit_what_we_together_do.pdf.

Pettersson, K. (2014). Informational Models in Deontic Logic: A Comment on "Ifs and Oughts" by Kolodny and MacFarlane. *Filosofiska Notiser* 1: 91–103.

Regan, D. (1980). *Utilitarianism and Cooperation*. New York: Oxford University Press.

Rosen, G. (2008). Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility. *Journal of Philosophy* 105: 591–610.

Spencer, J. and Wells, I. (forthcoming). Why Take Both Boxes? *Philosophy and Phenomenological Research*.

Zimmermann, T. (2000). Free Choice Disjunction and Epistemic Possibility. *Natural Language Semantics* 8: 255–290.

Michael Shaffer
Department of Philosophy
CH 365
St. Cloud State University
720 4th Ave. South
St. Cloud, MN 56301
shaffermphil@hotmail.com