Economics and Philosophy, **31** (2015) 259–274 © Cambridge University Press
doi:10.1017/S0266267115000127 First published online 7 May 2015
journals.cambridge.org/eap

TRACTABLE FALSIFIABILITY

RONEN GRADWOHL*, **ERAN SHMAYA**†

Abstract: We propose to strengthen Popper's notion of falsifiability by adding the requirement that when an observation is inconsistent with a theory, there must be a 'short proof' of this inconsistency. We model the concept of a short proof using tools from computational complexity, and provide some examples of economic theories that are falsifiable in the usual sense but not with this additional requirement. We consider several variants of the definition of 'short proof' and several assumptions about the difficulty of computation, and study their different implications on the falsifiability of theories.

Keywords: Falsifiability, computational complexity.

1. INTRODUCTION

Popper's notion of falsifiability is a foundational concept in philosophy of science that has been influential in economic theory. According to Popper, a theory is scientific if something could potentially occur that would contradict the theory's assertions. That is, if a theory is false, then there is some observation that conflicts with a prediction of the theory. The scientific assertions of a theory are then those that, if wrong, can be demonstrated as such.

The concept of falsifiability is illustrated in the following quote of Albert Einstein: 'No amount of experimentation can ever prove me right; a single experiment can prove me wrong.' What does it mean for an experiment to prove Einstein wrong? Such an experiment produces an outcome that is incompatible with some prediction of Einstein's theory.

* Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. Email: r-gradwohl@kellogg.northwestern.edu. URL: http://www.kellogg.northwestern.edu/faculty/Gradwohl/index.html.

† School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 6997801, Israel and Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA. Email: erans@post.tau.ac.il. URL: http://www.kellogg.northwestern.edu/faculty/directory/shmaya_eran.aspx.

Note, however, that theories are rarely formulated as collections of predictions, but by a set of laws which possibly include some unspecified parameters, from which the predictions of the theory can be derived. Thus, to prove Einstein wrong, i.e. to falsify his theory, one has to provide an experimental outcome *and* an argument that shows that the theory predicts that this outcome should not have occurred.

In this paper we propose an additional desideratum to Popper's notion of falsifiability: If a theory is wrong, then there should be an experiment or an observation that demonstrates this incorrectness by means of a 'short' proof. Without this additional requirement, it could be that there exists a false theory and experiments whose conclusions conflict with the theory, but such that these conclusions cannot be shown to contradict the theory in a reasonably short amount of time. We call a theory that satisfies Popper's notion and our additional requirement *tractably falsifiable*.

A good analogy to bear in mind is that of a court of law. Imagine a plaintiff claiming that some theory is wrong. The burden of proof is on the plaintiff, and he must produce some body of evidence against the theory. This evidence includes observations from the world and an argument that these observations contradict the theory. The court, in turn, has a protocol specifying the procedure for viewing evidence and examining arguments, after which it issues a verdict. Our main point is that the court is bounded in the amount of time it can allocate to the case. Therefore, any protocol specifying procedures must terminate in a 'short' amount of time: We rule out a situation in which the plaintiff's argument that the evidence contradicts the theory and the court's evaluation of this argument will only terminate after the universe no longer exists.

To model the concept of a protocol that terminates in a short amount of time we appeal to the tools of computational complexity. This discipline establishes a distinction between short and long protocols and arguments, formalized by polynomial-time and super-polynomial-time. At the core of computational complexity theory are numerous widely believed but yet unproven assumptions about the difficulty of computation, the most celebrated of which is the conjecture that $\mathcal{P} \neq \mathcal{NP}$.

As mentioned, our starting point is that both the protocol and the plaintiff's argument should be short. In the formalism of computational complexity, a short argument is of size that is polynomial in the amount of evidence. A short protocol is given by a polynomial-time algorithm that receives as input some observations and an argument by the plaintiff claiming that these observations conflict with the theory, and outputs a verdict. Importantly, the notion of polynomial-time, as well as all other notions from computational complexity, are asymptotic in nature. In this formulation, a protocol runs in polynomial-time if there exists some polynomial p such that the protocol decides whether a given evidence

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

violates the theory in time at most $p(n)$, where n is the amount of evidence.

Choice theory provides several examples of what we have in mind. Consider first a theory that asserts that an agent's preferences can be rationalized by a linear order over alternatives – that is, when choosing between a set of alternatives he picks the one that is ranked highest in his order. At first glance, if some observed choice behaviour violates this theory, then this violation can be easily demonstrated – the demonstration would simply consist of a list of all linear orders, along with a demonstration that none would have led to the observed choice behaviour. However, this proof is not 'short' – a full listing of all such orders is too long. Fortunately, there is a better way to prove the inconsistency of the theory, given by the strong axiom of revealed preference: A violation is given by a cycle in the agent's choice behaviour. Such a cycle proves the inconsistency, and does so in a concise way. The fact that the strong axiom of revealed preference characterizes choice functions that are consistent with the theory renders the theory tractably falsifiable in our sense: If the theory is not true – that is, if the agent's choices cannot be rationalized by a linear order – then there must be an observation (namely, a collection of choices) that violates the theory and a short argument (namely, a cycle) that a plaintiff can present to demonstrate this violation.

A contrasting example from choice theory is given by a model of choice by multiple rationales due to Kalai *et al.* (2002). According to this theory, an agent is endowed with multiple linear orders (say, two), and when confronted by a set of alternatives he chooses one that is highest ranked in either of his orders. Once again, observed choice behaviour can in principle be demonstrated to be in violation of this theory by enumerating all pairs of linear orders, but this demonstration will take much too long. Unlike the previous example, however, in this case there is no alternative short way to demonstrate this violation whenever it occurs.

As we shall see, this point is closely related, though not equivalent, to the fact, proved by Demuynck (2011), that it is an $\mathcal{NP}$-complete problem to determine whether choice behaviour is rationalizable by two linear orders. Similar proofs of $\mathcal{NP}$-completeness of rationalization have been given for other economic theories – see Kalyanaraman and Umans (2008, 2009), Apesteguia and Ballester (2010) and Demuynck (2014). In general, economists have been sceptical about the implication and meaning of these results, especially because of the worst-case assumption behind computational complexity analysis. Our focus is somewhat different than of the aforementioned papers: We are not concerned with the difficulty of rationalizing per se, but only with its implication about the falsifiability of the theory. What matters for us is whether there exists a short way to

demonstrate that a theory is wrong. We view this as a desirable, though idealistic, property of a scientific theory.

Our definition of a theory identifies a theory with its predictions. This approach, which is already strongly influenced by Popper's falsifiability stipulation, takes observable objects as primitive. Different theories make different predictions about relationships between these same observables, as in the examples of rational choice and choice by multiple rationales mentioned above. Economic theories fit this framework well, since the observables, such as choices and prices, have (or at least are assumed to have by economic theorists) a well-defined meaning independent of the economic theory. In contrast, this approach is not suitable for formalizing physical theories since 'physical observables' such as weight, time, electric current are already *theory laden*: they change their meaning from one physical theory to another, a point which is central in Pierre Duhem's (1954) argument.

Our paper is related to a recent paper of Chambers *et al.* (2014). They call a theory 'completely falsifiable' if, whenever the theory is incorrect, there exists some data set that falsifies it. Chambers *et al.* (2014) emphasize the requirement that the data set that falsifies a theory be finite. A typical example is the theory that the choice of an agent can be rationalized by a real-valued utility function, as opposed to a linear order. In this case, if the underlying set of alternatives is infinite, it may be that observed preference will admit no such rationalization but no finite data set will reveal this. Our notion of tractable falsifiability is more restrictive: Not only do we require that the data set that falsifies a theory be finite, we also require that the plaintiff be able to prove the falsification quickly. Consider again the theory that an agent's choice can be rationalized by two linear orders. This theory is completely falsifiable in the sense of Chambers *et al.* (2014): Even if the domain of alternatives is infinite, for every choice function that cannot be rationalized by a pair of linear orders over alternatives there is always a finite set of choices that cannot be rationalized. However, as we already mentioned, this theory is not tractably falsifiable, since it may be the case that there is no short proof that this data set has no rationalization: The plaintiff can produce evidence (a finite data set), but he has no quick way to show that this evidence is indeed inconsistent with the theory's predictions.

Our paper is also somewhat related to the paper of Echenique *et al.* (2011), which also involves falsifiability and computational complexity, but in a very different way. Echenique *et al.* (2011) study the question of whether computational considerations affect the scientific assertions of various theories. Their definition of 'scientific assertions,' however, is the one derived from the standard notion of falsifiability, without any computational considerations. The interface between falsifiability and computational complexity also plays a role in Chambers *et al.* (2011).

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

They study general revealed preference theories that are axiomatized by an existential quantifier over some unobserved relations. In particular, they argue that the problem of deciding whether a set of observations is consistent with the theory is in $\mathcal{NP}$.

There are other allusions in the literature for the relevance of computational complexity considerations to falsifiability. For example, Demuynck (2011) writes, 'From an empirical point of view, the fact that the verification of a certain choice model is $\mathcal{NP}$-complete shows that empirical refutation or acceptance of these models might be extremely difficult.' Our contribution is to provide a formal framework to analyse such assertions and to pin down the complexity assumptions on which they rely.

The relationship between falsifiability and computability was also suggested in the economic literature about testing stochastic forecasts. In that set-up an expert proposes a theory about some data generating process and this theory is put to test using the observed realization of the process. The literature studies manipulation of the test by strategic experts. Olszewski and Sandroni (2011) emphasize the implications to falsifiability of stochastic theories. Hu and Shmaya (2012) add the requirement that the test will be computable, which means that if some observed realization is inconsistent with the theory according to the test then there will be some finite proof for this inconsistency. Fortnow and Vohra (2009) require the test to be polynomial, which, in light of Hu and Shmaya's argument and the argument of this paper means that there exists a short proof that demonstrates the inconsistency.

We consider several possible approaches for proving that evidence falsifies a theory and several possible assumptions about the difficulty of computation, and study their implications for the notion of falsifiability. In section 2 we formalize the concepts of theory and evidence in a way that make them susceptible to complexity analysis. In section 3 we give our first definitions of argument and court's protocol and the derived notion of tractable falsifiability, and provide examples of theories that are not tractably falsifiable under the assumption that $\mathcal{NP} \neq co\mathcal{NP}$ (the class $co\mathcal{NP}$ is the complement of $\mathcal{NP}$, and the statement $\mathcal{NP} \neq co\mathcal{NP}$, while stronger than the famous $\mathcal{P} \neq \mathcal{NP}$, is still widely believed). In section 4 we consider a more general definition of a proof, in which the court is allowed to challenge the plaintiff with questions, and show that with this definition more theories become tractably falsifiable. In section 5 we then study the implication of the assumption that the universe is limited in the complexity of objects it can produce, so that we are only interested in falsifications of theories relative to this assumption. The appendix contains some formal definitions from computational complexity that we use. Our approach and examples should be intelligible without the definitions in the appendix, but we appeal to them in our proofs and in

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

places where the reference might be useful for readers who are familiar with computational complexity.

2. A FORMAL MODEL FOR THEORIES

Definition 2.1 (theory) A *theory* is a subset $T \subseteq \{0,1\}^*$, where $\{0,1\}^*$ is the set of all binary strings.

Consider the following simple example.

Example 2.2 A *society* is a graph whose nodes are called *individuals*, and in which two individuals are connected by an edge if they *know* each other. Every society can be encoded as an element of $\{0,1\}^*$ by the adjacency matrix of its graph.[1] The theory T_{cl} is the set of all elements in $\{0,1\}^*$ that represent a graph of some size n with a clique of size $\log n$.

This simple example illustrates our formalization of a theory. There is some set of conceivable objects (in this case conceivable societies, which may or may not have a large clique), and each object can be encoded as a binary string. A theory, modelled as a set of codes, says that only some of these objects (i.e. those whose codes are elements of the theory) can actually occur in our world. The description of the theory is usually not given by a list of the codes of these objects (which is usually infinite) but as a general property that all these objects satisfy. The theory T_{cl} says that in every society that occurs in our world there is a large set of individuals who know each other.

The coding of objects is necessary for complexity considerations, which are asymptotic in the length of the code. There is a strong sense in which all natural codes (i.e. codes that are not overly redundant and are simple to encode and decode) lead to the same computational consequences and are therefore equivalent for our purposes. For more on this issue, see for example Section 1.2.1 of Goldreich (2008). In the sequel we usually do not explicitly specify the code, and, slightly abusing terminology, identify an object with its code. (Strictly speaking, it is also possible that some string $x \in \{0,1\}^*$ is not a proper encoding of an object. For instance, in Example 2.2 it could be that $x \notin T_{cl}$ because $|x| \neq n^2$ for any n. For our purposes, however, this will not matter, since we assume that the properness of encodings is simple to check.)

We now return to the examples from the introduction: the theory of rational choice and the multiple rationales theory of Kalai *et al.* (2002).

[1] The adjacency matrix of an n-node graph is an $n \times n$ matrix in which the ijth entry is 1 if node i is connected to node j, and 0 otherwise.

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

Example 2.3 The *observed choice behaviour* of an individual is a collection of pairs (S, a), where S is a set of alternatives represented by a subset of $\{1, \ldots, m\}$ for some m, and a is the element of S that is the individual's choice. Every observed choice behaviour can be encoded as an element of $\{0,1\}^*$. The theory T_{RC} is the set of all elements $C \in \{0,1\}^*$ that represent observed choice behaviour satisfying the following condition: There exists a linear order over $\{1, \ldots, m\}$ for some m such that for every set of alternatives S in the collection C, the corresponding choice a is maximal in S for that linear order.

Example 2.4 The theory T_{KRS} is the set of all elements $C \in \{0,1\}^*$ that represent observed choice behaviour of an individual satisfying the following condition: There exists a *pair* of linear orders over $\{1, \ldots, m\}$ for some m such that for every set of alternatives S in the collection C, the corresponding choice a is maximal in S for at least one of the two linear orders.

If x is an object such that $x \notin T$ then the theory T predicts that the object x will never occur in our world. Thus, our definition of a theory identifies a theory with its predictions.

3. TRACTABLY FALSIFIABLE THEORIES

In principle, existence of an object that is not in the theory is a falsification of the theory: For the theory T_{cl} such a falsification will be a society that doesn't admit a large clique; for the theory T_{KRS} such a falsification will be an individual's observed choice behaviour that cannot be rationalized by two linear orders. In Definition 3.1 below, we propose to add the requirement that the assertion that the object is not in the theory should be demonstrable.

More formally, consider some theory T and an object $x \in \{0,1\}^*$, and suppose that $x \notin T$. In order to prove that $x \notin T$, the plaintiff may provide an argument $y \in \{0,1\}^*$. Here, again, we identify the argument with its coding as a binary string. The court then follows a protocol, which is abstracted by an algorithm V, that operates on both x and y. If $V(x, y) = 1$ then the court concludes that indeed $x \notin T$. If, however, $x \in T$, then it should be the case that regardless of the argument y provided by the plaintiff, the protocol $V(x, y) = 0$. Since we want a short proof, we will require that the algorithm V terminate in a reasonable amount of time, and this is formalized by the requirement that V be a polynomial-time algorithm. Finally, since we want V to be polynomial in the observation x, but V takes both x and y as input, we require the length of the argument y to be at most polynomial in the length of x.

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

Consider for example the theory T_{cl} and an object x that is some observed choice behaviour. Suppose that $x \notin T_{\mathrm{RC}}$, meaning that the choice behaviour x falsifies the model of rational choice. In order to prove that $x \notin T_{\mathrm{RC}}$, the plaintiff may provide an argument y – in this case, the argument would be a subset of the pairs of sets and choices $y = \{(S_1, a_1), \ldots, (S_k, a_k)\}$ that are contained in the observed behaviour x and form a cycle. The court's protocol here would be the following. Given x and y, the algorithm V checks that the elements of y are indeed in x, then checks that all the a_is are distinct, and finally checks that y induces a cycle: Namely, that for each $i \in \{1, \ldots, k\}$ and $j = (i \sim \mathrm{mod} \sim k) + 1$ it holds that $a_j \in S_i$. If all these are satisfied, then $V(x, y) = 1$, and otherwise $V(x, y) = 0$.

Now, if $x \notin T_{\mathrm{RC}}$ then the plaintiff can always provide such a y, and then when $V(x, y) = 1$ the court can conclude that indeed $x \notin T_{\mathrm{RC}}$. If, however, $x \in T_{\mathrm{RC}}$, then regardless of the argument y provided by the plaintiff, the protocol $V(x, y)$ will output 0, since there will never be a cycle. Note that in this example, y is shorter than x (and so in particular it is polynomial in the length of x). Also, note that the algorithm V that makes the checks above runs in polynomial time.

Definition 3.1 (tractably falsifiable theories) A theory $T \subseteq \{0,1\}^*$ is *tractably falsifiable* if there exists a polynomial $p(n)$ and a polynomial-time algorithm V such that the following two conditions hold:

1. For every $x \notin T$ there exists a $y \in \{0,1\}^*$ of length at most $p(|x|)$ such that $V(x, y) = 1$.
2. For every $x \in T$ and every $y \in \{0,1\}^*$ it holds that $V(x, y) = 0$.

A pair (x, y) as in the first bullet is called a *falsification*, where x is the *evidence* and y is the *argument* for $x \notin T$. The algorithm V is called the *falsification protocol* for T.

The first requirement in Definition 3.1 states the *completeness* of the falsification protocol: if the theory is wrong, i.e. if an object x such that $x \notin T$ does pop up in our world, then when x is observed there is an argument, or a way to demonstrate the fact that x refutes the theory. The second requirement states the *soundness*: the plaintiff cannot convince the court that a theory is wrong if the evidence x does not contradict the theory.

Definition 3.1 is essentially the definition of the complexity class $co\mathcal{NP}$, and we thus propose to identify the class of tractably falsifiable theories with this class. $co\mathcal{NP}$ is the complement of the class $\mathcal{NP}$ – see Appendix 1 for a formal definition. As we discussed in the Introduction, we think about the argument y against x in Definition 3.1 as a demonstration that a plaintiff can present in court to show that x

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

violates the theory. The demonstration must be short, but we make no assumptions about the complexity of finding it.

We now return to the examples from the beginning of the section:

Example 3.2 Assume $\mathcal{NP} \neq co\mathcal{NP}$. Then the theories T_{cl} and T_{KRS} from Examples 2.2 and 2.4 are not tractably falsifiable.

Thus, if $\mathcal{NP} \neq co\mathcal{NP}$ it may be the case that the theories are violated in our world, i.e. that there exists a society without a large clique and individuals whose observed choice behaviour cannot be rationalized by two linear orders, but still there is no short way to demonstrate the violations of the theories. We emphasize that these assertions do not rely solely on the assumption that $\mathcal{P} \neq \mathcal{NP}$, but rather on the stronger assumption that $\mathcal{NP} \neq co\mathcal{NP}$.

The statement of Example 3.2 follows from the facts that T_{cl} and T_{KRS} are $\mathcal{NP}$-complete (see Demuynck (2011) for a proof of the latter) and from Observation 3.3 below.

Observation 3.3 *Assume that $\mathcal{NP} \neq co\mathcal{NP}$. If a theory T is $\mathcal{NP}$-complete then T is not tractably falsifiable.*

This observation follows from our definition of tractable falsifiability as a theory in the complexity class $co\mathcal{NP}$ and the fact that if $\mathcal{NP} \neq co\mathcal{NP}$ then no theory that is $\mathcal{NP}$-complete can be in $co\mathcal{NP}$ (see for example Section 2.4.3 of Goldreich (2008) for a proof).

Note that if a theory is in $\mathcal{P}$, then it is tractably falsifiable, since $\mathcal{P} \subseteq co\mathcal{NP}$. Note also that even if a theory $T \notin \mathcal{P}$, this does not immediately imply that T is not tractably falsifiable – in particular, it could be the case that $T \in co\mathcal{NP} \setminus \mathcal{P}$.

It is also interesting to note that in principle there is no direct implication between the predictive power of a theory and the difficulty of demonstrating that a given outcome violates the theory. For a pair of theories T and T' let us say that T' has more predictive power than T if $T \subseteq T'$ (i.e. if any observation that is compatible with T is also compatible with T'). The theory T_{RC} is a sub-theory of T_{KRS}, and, as we have argued before, the former is tractably falsifiable while the latter is not. On the other hand, consider again the framework of Example 2.2 and let T' be the theory that in every society there is a pair of individuals that know each other. Clearly T_{cl} is a sub-theory T', but the former is not tractably falsifiable while the latter is.

The last example also shows that our definition of tractable falsifiability is demanding in the sense that we require the existence of a short demonstration for *every* object that violates the theory. So, even if the theory is not tractably falsifiable, it may be that some of its predictions can

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

be easily checked. For example, the theory T_{cl} implies that in every society there is at least one pair of people who knows each other. This prediction is easy to check, and a demonstration that a society includes no such pair will a fortiori demonstrate violation of T_{cl}.[2]

4. INTERACTION

Under the analogy of a court of law, in order to falsify a theory the plaintiff presents evidence and a short argument claiming that the evidence violates a theory, and the court bases its decision on the evidence and argument. In this section we relax the definition of tractable falsifiability by allowing for interaction between the plaintiff and the court – that is, we allow the court to present questions to the plaintiff, for which he must give answers as part of the protocol. The court would then base its decision on both the evidence and the plaintiff's answers to the court's questions. The possibility of interaction enlarges the set of theories that are falsifiable in a short amount of time. Interaction also might provide a way to think about academic dialogue and peer review. For example, suppose some researcher were to produce evidence that he claims is incompatible with some prediction of Einstein's theory. In addition to presenting this evidence, he may also be challenged by his peers and will have to provide an argument to defend his claim. If the researcher successfully counters these challenges this could count as strong evidence that his claim against Einstein's theory is valid. This is the essence of interaction.

With interaction, there is a need to specify the protocol for both the court and the plaintiff, which we model using a *two-party protocol*. A two-party protocol consists of two alternating algorithms that map a common input x and the history of messages thus far to a next message. More formally, what we mean by interaction of an algorithm V with an algorithm P on common input x is a sequence of messages $x_0, y_0, x_1, x_2, \ldots, x_k, y_k$ such that

$$\begin{aligned} x_0 &= x, \\ y_i &= V\left(x_0, y_0, x_1, y_1, \ldots, x_{i-1}\right), \text{and} \\ x_i &= P\left(x_0, y_0, \ldots, x_{i-1}, y_{i-1}\right). \end{aligned}$$

In the following we will also require that V run in polynomial time, and this will include the requirements that k and the length of all messages are polynomial in $|x|$. The output of the interaction is y_k.

[2] A similarly demanding approach was taken by Chambers *et al.* (2011). They define a theory to be completely falsifiable if *every* instance that violates the theory includes some finite data set that reflects this violation.

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

Definition 4.1 (interactively falsifiable theories) A theory $T \subseteq \{0,1\}^*$ is *interactively falsifiable* if for every $\varepsilon > 0$ there exists a two-party protocol with an algorithm P and a probabilistic polynomial-time algorithm V such that the following two conditions hold:

1. For every $x \notin T$, the algorithm V outputs 1 with probability at least $1 - \varepsilon$ after interacting with P on common input x.
2. For every $x \in T$ and algorithm P^*, V outputs 0 with probability at least $1 - \varepsilon$ after interacting with P^* on common input x.

The algorithms V is called the *interactive falsification protocol* for T. The algorithm P is called the *argument* or of the plaintiff.

The first requirement of Definition 4.1 states the completeness of the falsification protocol: if x is indeed a falsification of T then the plaintiff has a way to convince the court that T is wrong – namely, he follows the procedure P. The second requirement states soundness: If x does not falsify the theory T, then regardless of the argument P^* made by the plaintiff the court will not be convinced otherwise. Note that while the court is limited to a polynomial time algorithm V, the plaintiff is not restricted. Definition 4.1 is very close to the definition of interactive proofs, initially introduced by Goldwasser *et al.* (1989) and Babai (1985), and characterized by the complexity class $\mathcal{IP}$ (see Appendix).

To see the power of interaction, contrast the following with Example 3.2.

Example 4.2 The theories T_{cl} and T_{KRS} from Examples 2.2 and 2.4 are interactively falsifiable.

This holds because of the following observation.

Observation 4.3 *If T is in $\mathcal{NP}$ then T is interactively falsifiable.*

Many economic theories, particularly those claiming that behaviour can be rationalized, are in the class $\mathcal{NP}$ (see also Chambers *et al.* 2011). Observation 4.3 then implies that these theories can be interactively falsified. The proof of Observation 4.3 makes use of a deep theorem in complexity theory proved by Shamir (1992), and we therefore defer the proof to the Appendix.

5. TRACTABLE UNIVERSE

The definitions of tractable falsifiability and interactive falsifiability are 'worst-case' definitions, in the sense that the falsification protocol must be short for *any* object that contradicts a given theory. In this section we add another twist to our framework and restrict our attention to

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

objects that 'typically' appear in the universe, and study the implication of this restriction on falsifiability. Of course, an immediate difficulty is, what is typical? This question is addressed by the theory of average-case complexity initiated by Levin (1986).

Levin's theory begins with the Church–Turing thesis. The Church–Turing thesis was formulated in the beginning of the 20 century following the work of various authors, most notably Alan Turing and Alonzo Church, to formally define the intuitive notion of effective computability. In its most simple form the hypothesis identifies effective computability with Turing Machine computability. A more extreme version states that the outcome of any physical process is a Turing computable function possibly with appeal to pure randomness. There is an extensive literature in philosophy about the ramification and implication of the physical Church–Turing thesis. The Church–Turing thesis also appears in economic literature. For example it is natural to assume that the choice of an economic agent, as the outcome of a physical or cognitive process, is a computable function of the given set of choices.

Intuitively, the physical Church–Turing thesis implies that the universe can be simulated by a computer with unlimited time and memory resources and access to pure randomness. Levin's theory relies on a stronger version of the hypothesis, sometimes called the *Strong Church–Turing thesis*, that the objects that appear in the universe can be created in polynomial-time (in the size of their canonical description). We call a universe that satisfies this assumption *tractable*.[3]

Recall that in order for a theory to be tractably falsifiable (Definition 3.1), for every instance x that violates the theory there must be a way to demonstrate this violation. In the following we modify the definition by requiring an argument only against 'typical' instances, which are those that can be produced in a tractable universe. However, because we also assume some randomness in the universe, we need to allow for a small probability that the universe does end up producing an instance that violates the theory, even though this violation cannot be demonstrated.

We follow Levin's theory of average-case complexity, which analyses the complexity of a decision problem according to the time it takes to answer it over instances that are drawn from a distribution that can be sampled in polynomial time. The following definition is very similar to Definition 3.1, except that the universe is restricted to produce objects using some probabilistic polynomial-time algorithm A.

[3] The Strong Church–Turing thesis is challenged by quantum computers, which seems to be physically realizable in principle but cannot be simulated efficiently on a Turing Machine (Bernstein and Vazirani 1997).

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

Definition 5.1 (tractably falsifiable theories in a tractable universe) A theory $T \subseteq \{0,1\}^*$ is *tractably falsifiable in a tractable universe* if for every probabilistic polynomial-time algorithm A and every $\varepsilon > 0$ there exists a polynomial $p(n)$ and a polynomial-time algorithm V such that the following two conditions hold:

1. If $x \in T$ then $V(x,y) = 0$ for every $y \in \{0,1\}^*$.
2. If $x = A(0^n)$ then

$$\mathcal{P}\big(x \notin T \text{ and } V(x,y) = 0 \text{ for all } y \in \{0,1\}^* \text{ such that } |y| \leq p(n)\big) < \varepsilon,$$

where the probability is over the internal randomization of A.

While in standard (worst-case) complexity theory there are various conjectures about the difficulty of problems, many of which are strongly believed to be true, the world of average-case complexity is significantly more complex. Impagliazzo (1995) considers various scenarios that are possible in terms of the average-case difficulty of problems, and calls each a possible *world*. For our purposes, consider the world he calls *Heuristica*. In this world problems in $\mathcal{NP}$ might be intractable in the worst-case, but are still tractable on average when the input is randomly drawn from a polynomial-time algorithm. Thus, for every probabilistic polynomial-time algorithm A as in Definition 5.1 there exists an algorithm that determines whether an output of A is in T with running time which is, with high probability, polynomial.

With the weaker definition of tractable falsifiability stated in Definition 5.1 we get the following example, which once again stands in contrast to Example 3.2.

Example 5.2 In Heuristica, the theories T_{cl} and T_{KRS} from Examples 2.2 and 2.4 *are* tractably falsifiable in a tractable universe.

This holds because of the following observation:

Observation 5.3 *In Heuristica, every theory that is in $\mathcal{NP}$ is tractably falsifiable in a tractable universe.*

Proof: Fix A and ε as in Definition 5.1, and let q be a polynomial and B be an algorithm for deciding membership T such that $\mathcal{P}\big(\mathrm{TIME}(B,x) > q(n)\big) < \varepsilon$ for $x = A(0^n)$, where $\mathrm{TIME}(B,x)$ is the running time of B on x. Such an algorithm B exists in Heuristica. Let $V(x,y) = B(x)$ if $\mathrm{TIME}(B,x) \leq q(n)$ and $V(x,y) = 0$ otherwise. ■

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

REFERENCES

Apesteguia, J. and M. Ballester. 2010. The computational complexity of rationalizing behavior. *Journal of Mathematical Economics* 46: 356–363.

Babai, L. 1985. Trading group theory for randomness. In *Proceedings of the 17th ACM Symposium on the Theory of Computing* (STOC '85), Providence, RI, 421–429. New York, NY: Association for Computing Machinery.

Bernstein, E. and U. Vazirani. 1997. Quantum complexity theory. *SIAM Journal on Computing* 26: 1411–1473.

Chambers, C., F. Echenique and E. Shmaya. 2011. General revealed preference theory. Caltech SS Working Paper 1332.

Chambers, C., F. Echenique and E. Shmaya. 2014. The axiomatic structure of empirical content. *American Economic Review* 104: 2303–2319.

Demuynck, T. 2011. The computational complexity of boundedly rational choice behavior. *Journal of Mathematical Economics* 47: 425–433.

Demuynck, T. 2014. The computational complexity of rationalizing Pareto optimal choice behavior. *Social Choice and Welfare* 42: 529–549.

Duhem, P. 1954. *The Aim and Structure of Physical Theory*. Princeton, NJ: Princeton University Press.

Echenique, F., D. Golovin and A. Wierman. 2011. A revealed preference approach to computational complexity in economics. In *Proceedings of the 12th ACM Conference on Electronic Conference* (EC '11), San Jose, CA, 101–110. New York, NY: Association for Computing Machinery.

Fortnow, L. and R. V. Vohra. 2009. The complexity of forecast testing. *Econometrica* 77: 93–105.

Goldreich, O. 2008. *Computational Complexity: A Conceptual Perspective*. New York, NY: Cambridge University Press.

Goldwasser, S., S. Micali and C. Rako. 1989. The knowledge complexity of interactive proof systems. *SIAM Journal on Computing* 18: 186–208.

Hu, T. and E. Shmaya. 2012. Expressible inspections. *Theoretical Economics* 8: 263–280.

Impagliazzo, R. 1995. A personal view of average-case complexity. In *Proceedings of the Tenth Annual IEEE Structure in Complexity Theory Conference*, Minneapolis, MN, 134–147. Los Alamitos, CA: IEEE Computer Science.

Kalai, G., A. Rubinstein and R. Spiegler. 2002. Rationalizing choice functions by multiple rationales. *Econometrica* 70: 2481–2488.

Kalyanaraman, S. and C. Umans. 2008. The complexity of rationalizing matchings. *Algorithms and Computation: 19th International Symposium, ISAAC 2008, Gold Coast, Australia, December 2008, Proceedings*, Lecture Notes in Computer Science 5369, ed. S.-H. Hong, H. Nagamochi and T. Fukunaga, 171–182. Berlin: Springer.

Kalyanaraman, S. and C. Umans. 2009. The complexity of rationalizing network formation. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science* (FOCS 2009), Atlanta, GA, 485–494. Los Alamitos, CA: IEEE Computer Science.

Levin, L. A. 1986. Average case complete problems. *SIAM Journal on Computing* 15: 285–286.

Olszewski, W. and A. Sandroni. 2011. Falsiability. *American Economic Review* 101: 788–818.

Shamir, A. 1992. IP = PSPACE. *Journal of the ACM* 39: 869–877.

APPENDIX

1. Definitions from Computational Complexity

In this section we briefly and somewhat informally provide the computational complexity background used in this paper. See Goldreich (2008) or any other textbook on computational complexity for further details.

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

While in this paper we study theories, as defined in Definition 2.1, the standard nomenclature for the same object in the field of computational complexity is called a *decision problem*. The corresponding problem is to decide, for a given input $x \in \{0,1\}^*$ whether $x \in T$. For every code $x \in \{0,1\}^*$ we denote its length by $|x|$. An algorithm A is called *polynomial-time algorithm* if $\text{Time}(B,x) < p(|x|)$ for every $x \in \{0,1\}^*$ where $\text{Time}(B,x)$ is the running time (more specifically, number of operations) of B on x.

A decision problem T is in *the class* $\mathcal{P}$ if there is a polynomial-time algorithm A such that that, for any input $x \in \{0,1\}^*$, $A(x) = 1$ if and only if $x \in T$.

A decision problem T is in *the class* $\mathcal{NP}$ if there exists a polynomial $p(n)$ and a polynomial-time algorithm V such that the following two conditions hold:

1. For every $x \in T$ there exists a $y \in \{0,1\}^*$ of length at most $p(|x|)$ such that $V(x,y) = 1$.
2. For every $x \notin T$ and every $y \in \{0,1\}^*$ it holds that $V(x,y) = 0$.

Roughly, this is the class for which there is a polynomial-time algorithm that verifies that $x \in T$ with the aid of a *witness* y. Clearly $\mathcal{P} \subseteq \mathcal{NP}$. The widely believed assumption that $\mathcal{P} = \mathcal{NP}$ is at the core of computational complexity theory.

The hardest problems in $\mathcal{NP}$ are called $\mathcal{NP}$-complete. We don't provide the formal definition here since we don't make use of the concept of $\mathcal{NP}$-completentess in our definitions, but we make use of the fact that the decisions problems in Examples 2.2 and 2.4 are $\mathcal{NP}$-complete.

The class $co\mathcal{NP}$ is the class of complements of decision problems in $\mathcal{NP}$ – it contains those decision problems $\{\{0,1\}^* \setminus T : T \in \mathcal{NP}\}$. Definition 3.1 characterizes this class. It is widely believed that $\mathcal{NP} \neq co\mathcal{NP}$. This assumption is stronger than $\mathcal{P} \neq \mathcal{NP}$. (Indeed, if $\mathcal{P} = \mathcal{NP}$ then $\mathcal{P} = \mathcal{NP} = co\mathcal{NP}$ since the class $\mathcal{P}$ is closed under complement.)

2. Proof of Observation 4.3

The proof makes use of a non-trivial theorem of complexity theory. Before describing the theorem we define two additional complexity classes: $\mathcal{PSPACE}$ and $\mathcal{IP}$.

The class $\mathcal{PSPACE}$ contains all decision problems T for which there is an algorithm that, on any input x, determines whether or not $x \in T$, and such that the following holds: The number of bits of *memory* used by the algorithm must be at most polynomial in $|x|$. It is easy to verify that $\mathcal{NP} \subseteq \mathcal{PSPACE}$.

A decision problem $T \subseteq \{0,1\}^*$ is in the class $\mathcal{IP}$ if for every $\varepsilon > 0$ there exists a two-party protocol with an algorithm P and a probabilistic polynomial-time algorithm V such that the following two conditions hold:

1. For every $x \in T$, the algorithm V outputs 1 with probability at least $1 - \varepsilon$ after interacting with P on common input x.
2. For every $x \notin T$ and algorithm P^*, V outputs 0 with probability at least $1 - \varepsilon$ after interacting with P^* on common input x.

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press

Shamir (1993) proved that the classes $\mathcal{IP}$ and $\mathcal{PSPACE}$ are identical. We use Shamir's theorem in the proof of Observation 4.3.

Proof of Observation 4.3: According to definition 4.1, a theory T is interactively falsifiable if and only if its complement is in $\mathcal{IP}$. Indeed, an interactive proof that $x \notin T$ can be viewed as an interactive proof that $x \in T^c$, and vice versa. Shamir (1993) proved that the classes $\mathcal{IP}$ and $\mathcal{PSPACE}$ are equal. Since $\mathcal{PSPACE}$ is closed under complement, it follows that a theory is interactively falsifiable if and only if it is in $\mathcal{PSPACE}$. The observation follows from the fact that $\mathcal{NP} \subseteq \mathcal{PSPACE}$. ■

BIOGRAPHICAL INFORMATION

Ronen Gradwohl is an Assistant Professor of Managerial Economics and Decision Sciences at the Kellogg School of Management, Northwestern University. His current research focuses on the strategic effects of information concealment and disclosure, with particular emphasis on individual privacy and institutional transparency.
Eran Shmaya is a Senior Lecturer at the School of Mathematical Sciences, Tel Aviv University and an Assistant Professor at Northwestern University. His research is in game theory and mathematical economics.

https://doi.org/10.1017/S0266267115000127 Published online by Cambridge University Press