Bringing Order to the Cognitive Fallacy Zoo

Ardavan S. Nobandegani[1,3], **William Campoli**[2], **& Thomas R. Shultz**[2,3]
{ardavan.salehinobandegani, william.campoli}@mail.mcgill.ca
{thomas.shultz}@mcgill.ca
[1]Department of Electrical & Computer Engineering, McGill University
[2]School of Computer Science, McGill University
[3]Department of Psychology, McGill University

Abstract

Investigations into human reasoning, judgment and decision-making have led to the finding of numerous cognitive biases and fallacies, with new ones continually emerging, leading to a state of affairs which can fairly be characterized as the cognitive fallacy zoo! In this work, we formally present a principled way to bring order to this zoo. We introduce the idea of establishing implication relationships (IRs) between cognitive fallacies, formally characterizing how one fallacy implies another. IR is analogous to, and partly inspired by, the fundamental concept of reduction in computational complexity theory. We present several examples of IRs involving experimentally well-documented cognitive fallacies: base-rate neglect, availability bias, conjunction fallacy, decoy effect, framing effect, and Allais paradox. We conclude by discussing how our work: (i) allows for identifying those pivotal cognitive fallacies whose investigation would be the most rewarding research agenda, and importantly (ii) permits a systematized, guided research program on cognitive fallacies, motivating influential theoretical as well as experimental avenues of future research.

Keywords: Cognitive biases and fallacies; cognitive fallacy map; cognitive processes;

1 Introduction

Over the past few decades, empirical investigations into human reasoning, judgment and decision-making have led to the discovery of new cognitive fallacies, giving rise to a large, ever-growing number of documented biases and fallacies, a state of affairs which can fairly be characterized as a zoo[1] of cognitive fallacies (e.g., Tversky & Kahneman, 1973, 1981b). A glance at over a hundred cognitive fallacies listed on Wikipedia attests to this claim (see Fig. 1).

In this work, we formally present a principled way to bring order to the cognitive fallacy zoo, allowing for a precise characterization of how various fallacies relate to one another. We introduce the idea of establishing an *implication relationship* (IR) (denoted by $\rightsquigarrow$) between a pair of cognitive fallacies, formally characterizing how the occurrence of one fallacy implies another. More formally, for two cognitive fallacies A, B, the expression $A \rightsquigarrow B$ denotes that A leads to B, i.e., the occurrence of A logically implies the occurrence of B. As a proof-of-concept, we present several examples of IRs involving experimentally well-documented cognitive fallacies: base-rate neglect (Tversky & Kahneman, 1981a), availability bias (Kahneman & Tversky, 1973), conjunction fallacy (Kahneman & Tversky, 1983), decoy effect (Huber, Joel, & Puto, 1982), framing effect (Tversky & Kahneman, 1981b) and Allais paradox (Allais, 1953).

The idea of establishing IRs between pairs of cognitive fallacies is analogous to, and partly inspired by, the foundational concept of reduction in computational complexity theory (see Karp, 1972; Papadimitriou, 2003; Arora & Barak, 2009; Sipser 2006), which has played a profound role in theoretical computer science, allowing to formally establish how various computational problems relate to each other and how the solution to one sheds light on that of another. After a brief discussion on the role of reduction in computational complexity, we return to the formal characterization of the notion of IR and subsequently present several examples of IRs involving experimentally well-documented cognitive fallacies. But first, some historical notes on reduction in computational complexity.

2 Reduction in Computational Complexity

The notion of reduction plays a fundamental role in computational complexity theory, and in theoretical computer science more generally. Informally put, a computational problem A is *reducible* to computational problem B, if every instance of A can be transformed into an instance of B. Therefore, the reduction of A to B offers an indirect way of solving A, by first reducing A to B, and then solving B.

To further clarify the idea of reduction, we provide two examples. As a first example, consider two well-known computational problems, namely, HAMILTONIAN-PATH and HAMILTONIAN-CYCLE. The HAMILTONIAN-PATH problem is defined as follows: given a (directed) graph G, is there a path which visits every node of G exactly once? The HAMILTONIAN-CYCLE is defined as follows: given a (directed) graph G, is there a cycle which visits every node of G exactly once? It turns out that HAMILTONIAN-CYCLE is reducible to HAMILTONIAN-PATH. Given that the definitions HAMILTONIAN-CYCLE and HAMILTONIAN-PATH are closely related (since a cycle is a path with its endpoints coinciding), this reduction is not especially surprising.

As a second example, let us consider HAMILTONIAN-PATH together with the 3-COLORABILITY problem, defined as follows: given a graph G and 3 distinct colors, can you color the nodes of G such that the endpoints of every edge are colored differently? At fist glance, the HAMILTONIAN-PATH and 3-COLORABILITY appear to have no connection with one another whatsoever. Surprisingly, however, it

[1]The term 'cognitive fallacy zoo' is inspired by an analogous terminology in the computational complexity literature, called 'complexity zoo.' For details, visit: https://complexityzoo.uwaterloo.ca/Complexity_Zoo

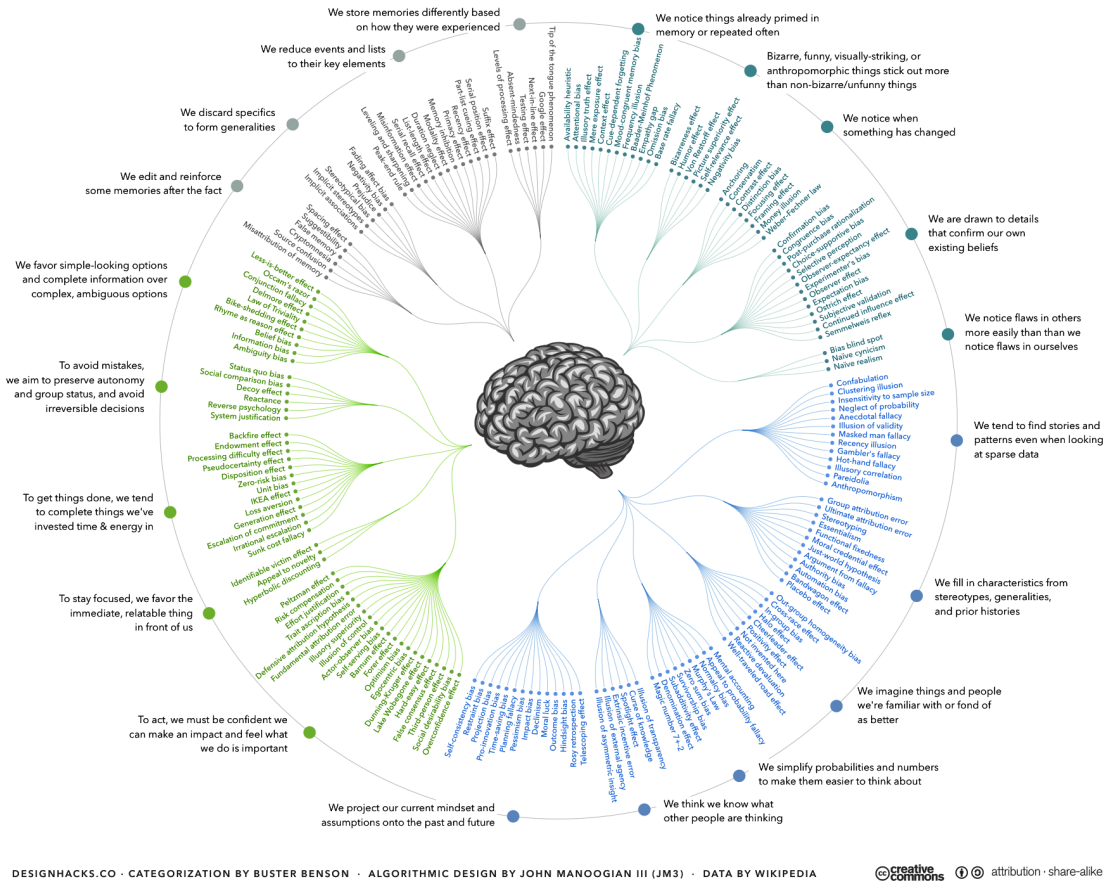

Figure 1: Cognitive fallacies listed on Wikipedia. Besides qualitatively categorizing them into classes depending on the context in which they occur, to date there exists no principled way of bringing order to these fallacies, allowing for formally characterizing how one fallacy relates to another.

turns out that HAMILTONIAN-PATH can be reduced to 3-COLORABILITY.[2] Thus, the question of whether a graph G has a Hamiltonian path can be resolved by answering if a corresponding graph G' is 3-colorable.

The idea of reduction has had profound implications for theoretical computer science, allowing for formally connecting seemingly unrelated computational problems to one another (see Fig. 2(a)). Had reduction not been introduced into theoretical computer science, every computational problem would have had to be investigated on its own, because the solution to one would not have shed any light on the solution to others. It was a major breakthrough when Karp (1972) showed that a key computational problem called SATISFIABILITY could be reduced to a number of other well-known computational problems, a contribution for which he was eventually awarded the Turing award in 1985. It is also worth noting that the (in)famous $\mathcal{P}$ vs. $\mathcal{NP}$ problem in theoretical computer science, at its core, concerns the possibility or impossibility of establishing a particular form of reduction.

[2]The reduction can be established by a chain of straightforward reductions: HAMILTONIAN-PATH to SAT, SAT to 3SAT, and finally, 3SAT to 3-COLORING.

One might wonder if an idea broadly analogous to reduction in computational complexity could be introduced into cognitive science, allowing for formally connecting seemingly different cognitive fallacies with one another, and, hence, bring order to the cognitive fallacy zoo in a principled manner. Primarily motivated by this, and, by analogy with the notion of reduction in theoretical computer science, we introduce the idea of establishing IRs between cognitive fallacies, formally characterizing how one fallacy would imply another.

3 Implication Relationships: Formalization

In what follows, we first formally introduce the idea of establishing an *implication relationship* (IR) (denoted by $\rightsquigarrow$) between a pair of cognitive fallacies, followed by several examples of IRs involving experimentally well-documented cognitive fallacies.

Definition (Implication Relationship). For two cognitive fallacies/biases A, B, the fallacy A is said to *implicate* the fallacy B (denoted by $A \rightsquigarrow B$) if and only if the occurrence of A *logically implies* the occurrence of B.

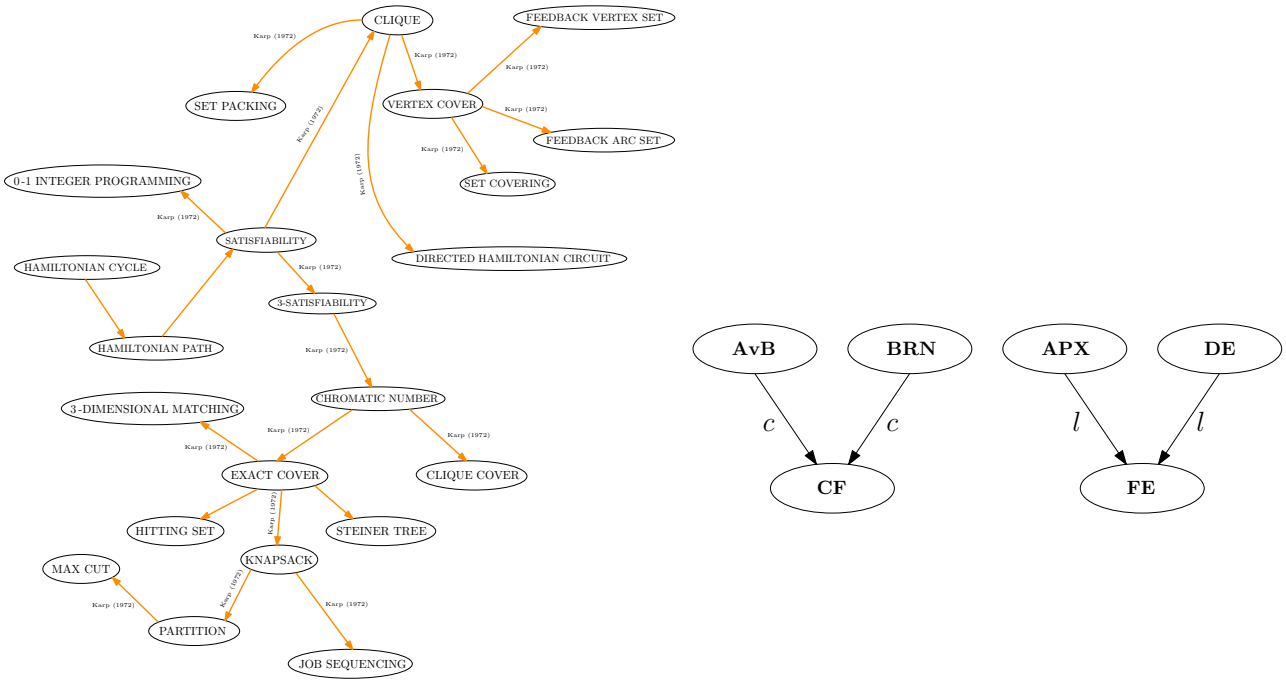

Figure 2: **(a)** A map showing reductions (directed gold lines) between a set of important computational problems (ovals) in theoretical computer science, formally characterizing how one problem is related to another. The first demonstration of a particular reduction from one problem to another is listed on the corresponding arrow between the two. **(b)** The IRs formally established in this paper, among well-known cognitive fallacies (AvB: Availability bias, BRN: base-rate neglect, CF: conjunction fallacy, APX: Allais paradox, DE: Decoy effect, FE: framing effect). Letters c (for causal) and l (for logical) on the arrows specify the type of an IR; see the Discussion section for details.

4 Examples on Implication Relationships

As a proof-of-concept, We next present several examples of IRs involving experimentally well-documented cognitive fallacies, namely, base-rate neglect (Tversky & Kahneman, 1981a), availability bias (Kahneman & Tversky, 1973), conjunction fallacy (Kahneman & Tversky, 1983), decoy effect (Huber et al., 1982), framing effect (Tversky & Kahneman, 1981b), and Allais paradox (Allais, 1953).

4.1 Case Study 1: Decoy Effect $\rightsquigarrow$ Framing Effect

As our first example, we formally establish an IR between two well-documented cognitive fallacies, namely, the decoy effect (DE) and the framing effect (FE).

The Framing Effect (FE): If people produce different responses for two equivalent tasks, the framing effect (FE) has occurred (Tversky & Kahneman, 1981b; Kahneman & Tversky, 1984). In that light, FE is a violation of the extensionality principle (Bourgeois-Gironde & Giraud, 2009), which prescribes that two equivalent tasks should elicit the same response.

FE is well captured by Tversky & Kahneman (1981b): Subjects were asked to "imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume the exact scientific estimate of the consequences of the programs are as follows." In one condition, subjects were presented with a choice between Programs A and B, while in another condition, subjects were asked to choose between Programs C and D:

Program A: 200 people will be saved.

Program B: There is a $\frac{1}{3}$ probability that 600 people will be saved, and a $\frac{2}{3}$ probability that no people will be saved.

Program C: 400 people will die.

Program D: there is a $\frac{1}{3}$ probability that nobody will die, and a $\frac{2}{3}$ probability that 600 people will die.

Despite the equivalence of these Programs pairs, a majority of the first group preferred Program A (= C), while a majority of the second group preferring Program D (= B).

The Decoy Effect (DE): The decoy effect (DE) refers to a change in people's preference between two options, when presented with a third *asymmetrically-dominated* option, i.e., an option which is inferior to one option in all respects, but, in comparison to the other option, it is inferior in some respects and superior in others. In that light, DE is a violation of the independence of irrelevant alternative axiom of rational choice theory (Ray, 1973), which prescribes the following: If A is preferred to B out of the choice set $\{A,B\}$, introducing a third option X, hence expanding the choice set to $\{A,B,X\}$, should not make B preferable to A.

We are now well-positioned to formally present our result.

Proposition 1. *The following holds:*

$$\text{DE} \rightsquigarrow \text{FE}.$$

Proof. According to normative principles, preference for the choice sets $\{A,B\}$ and $\{A,B,X\}$ should be the same, with X being an asymmetrically-dominated option. The rationale is the following: Since X is inferior to one option in *all* respects, rationally X should never be chosen; hence, the preference pattern for the choice sets $\{A,B\}$ and $\{A,B,X\}$ should be identical. Therefore, whenever people's preference pattern for the choice sets $\{A,B\}$ and $\{A,B,X\}$ differs (which is the case for DE), it logically implies the violation of the extensionality principle, hence granting the occurrence of FE. This concludes the proof. ■

The message of Proposition 1 is simple: From the standpoint of normative principles, the two choice sets $\{A,B\}$ and $\{A,B,X\}$ (with X being an asymmetrically-dominated option) are equivalent, therefore people's showing different preference patterns for the two choice sets, as is the case in DE, is a clear indication of FE. Proposition 1, therefore, formally establishes that the occurrence of DE leads to the occurrence of FE.

4.2 Case Study 2: Base-Rate Neglect $\rightsquigarrow$ Conjunction Fallacy

As our second example, we formally establish an IR between another pair of well-documented cognitive fallacies, namely, the base-rate neglect (BRN) and the conjunction fallacy (CF). BRN and CF can be characterized as follows.

The Base-Rate Neglect (BRN): Base-rate neglect (BRN) (Tversky & Kahneman, 1981a) refers to people not considering prior probabilities in their judgments under uncertainty.

The Conjunction Fallacy (CF): For two events A,B and presented with evidence e, people judge the probability of the event $A \cap B$ to be greater than that of A (or B), in isolation. That is, more formally, people judge: $\mathbb{P}(A \cap B|e) > \mathbb{P}(A|e)$. In that light, CF is a clear violation of the axioms of probability (since $\forall A,B,\ A \cap B \subseteq A \Rightarrow \mathbb{P}(A \cap B|e) \leq \mathbb{P}(A|e)\ \forall e \neq \varnothing$; that is, the probability of a subset of Y, in principle, cannot be greater than that of Y).

CF is well captured in the famous Linda experiment by Tversky and Kahneman (1981). Presented with a description (e) of Linda, a politically active, single, outspoken, and very bright 31-year-old female, people overwhelmingly judge that Linda is more likely to be a feminist bankteller ($A \cap B$) than to be a bankteller (A).

We are now well-positioned to formally present our result.

Proposition 2. *The following holds.*

$$\text{BRN} \rightsquigarrow \text{CF}.$$

Proof. Since $\mathbb{P}(A \cap B|e) = \mathbb{P}(e|A \cap B)\mathbb{P}(A \cap B)$ and $\mathbb{P}(A|e) = \mathbb{P}(e|A)\mathbb{P}(A)$, we have:

$$\frac{\mathbb{P}(A \cap B|e)}{\mathbb{P}(A|e)} = \frac{\mathbb{P}(e|A \cap B)}{\mathbb{P}(e|A)} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)},$$

where the term $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$ indicates the ratio between priors $\mathbb{P}(A \cap B)$ and $\mathbb{P}(A)$. If BRN occurs (which results in the term $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$ being dropped), it follows that:

$$\frac{\mathbb{P}(A \cap B|e)}{\mathbb{P}(A|e)} = \frac{\mathbb{P}(e|A \cap B)}{\mathbb{P}(e|A)}.$$

Assuming that $\mathbb{P}(e|A \cap B) > \mathbb{P}(e|A)$, which is the case in the context of CF (see the Linda experiment discussed above), it follows that:

$$\frac{\mathbb{P}(A \cap B|e)}{\mathbb{P}(A|e)} = \frac{\mathbb{P}(e|A \cap B)}{\mathbb{P}(e|A)} > 1 \Rightarrow \mathbb{P}(A \cap B|e) > \mathbb{P}(A|e),$$

hence CF occurs. This completes the proof. ■

In simple terms, Proposition 2 shows that the occurrence of BRN leads to the occurrence of CF.

4.3 Case Study 3: Allais Paradox $\rightsquigarrow$ Framing Effect

As our third example, we formally establish an IR between the Allais paradox (APX) and FE. APX can be characterized as follows. (See Sec. 4.1 for the characterization of FE.)

The Allais Paradox (APX): The Allais paradox refers to an observed reversal in participants' choices in two different experiments, each of which consists of a choice between two gambles, A and B, while in fact, according to the independence axiom of rational decision-making (Von Neumann & Morgenstern, 1953), no such a reversal should occur. That is, although the independence axiom grants the equivalence of the two experiments, the pattern of people's preference nevertheless reverses from the first experiment to the second.[3]

Proposition 3. *The following holds:*

$$\text{APX} \rightsquigarrow \text{FE}.$$

Proof. The proof is evident from the characterization of APX given above. Although the independence axiom of rational decision-making (Von Neumann & Morgenstern, 1953) grants the equivalence of the two experiments entertained in APX, the pattern of people's preference nevertheless reverses from one to the other. That is, in the case of APX, people produce different responses for two equivalent experiments. Therefore, the occurrence of the Allais paradox logically implies the occurrence of the framing effect. This concludes the proof. ■

4.4 Case Study 4: Availability Bias $\rightsquigarrow$ Conjunction Fallacy

As our final example, we formally establish an IR between the well-documented Availability bias (AvB) and CF. AvB can be concisely characterized as follows: (See Sec. 4.2 for the characterization of CF.)

The Availability Bias (AvB): Extreme events come to mind easily, people overestimate their probabilities, and overrepresent them in decision-making (Tversky & Kahneman,

[3]The reader is referred to Allais (1953) for a clear description of the two experiments.

1973; Lieder et al., 2018; Nobandegani et al., 2018). Formally, people overestimate the probability of an event o, $p(o)$, proportional to the absolute value of its subjective utility $u(o)$ (Lieder et al., 2018; Bordalo, Gennaioli, & Shleifer, 2012). That is, people's subjective probability of event o, $q(o)$, is given by[4] $q(o) \propto p(o)|u(o)|$.

Proposition 4. *Let o_1 and o_2 be two events, and let $o_\wedge$ denote the event corresponding to the occurrence of o_1 and o_2 together, i.e., the one corresponding to the conjunction of the two events o_1 and o_2. Assuming that $\forall i = 1,2,\ |u(o_\wedge)| \gg |u(o_i)|$, the following holds:*

$$\text{AvB} \rightsquigarrow \text{CF}.$$

Proof. According to the characterization of AvB given above, $q(o_\wedge) \propto p(o_\wedge)|u(o_\wedge)|$ and $\forall i = 1,2,\ q(o_i) \propto p(o_i)|u(o_i)|$. We have,

$$\forall i = 1,2, \quad \frac{q(o_\wedge)}{q(o_i)} = \frac{p(o_\wedge)}{p(o_i)} \frac{|u(o_\wedge)|}{|u(o_i)|}.$$

It follows from the axioms of probability that $\forall i = 1,2,\ p(o_\wedge) \leq p(o_i)$; hence, $\forall i = 1,2,\ \frac{p(o_\wedge)}{p(o_i)} \leq 1$. However, since $\forall i = 1,2,\ |u(o_\wedge)| \gg |u(o_i)|$, it follows that $\frac{|u(o_\wedge)|}{|u(o_i)|} \gg 1,\ \forall i = 1,2$. Therefore, altogether, $\forall i = 1,2,\ \frac{q(o_\wedge)}{q(o_i)} > 1$ which implies $\forall i = 1,2,\ q(o_\wedge) > q(o_i)$, granting the validity of the conjunction fallacy (CF). This concludes the proof. ■

The message of Proposition 4 is simple. If people judge the conjunction of two events to be much more extreme than each of them individually (i.e., $\forall i = 1,2,\ |u(o_\wedge)| \gg |u(o_i)|$), then the occurrence of AvB leads to the occurrence of CF.

5 General Discussion

In this work, we introduce the notion of implication relation (IR) between a pair of cognitive fallacies, formally characterizing how one would logically imply the other.

A crucial initial step in establishing IRs between cognitive fallacies is to provide a characterization of the cognitive fallacies involved in those IRs, i.e., to specify, for each cognitive fallacy, what instances and/or circumstances belong to the class of that cognitive fallacy. In Sec. 4, we first provide a characterization of the cognitive fallacies of interest, followed by formally establishing IRs. Particularly, we provide a broad characterization of the cognitive fallacies we are interested in, with those characterizations being primarily guided by experimental findings. As such, these characterizations could be arguably made more precise and/or broadened as future research deepens our understanding of the cognitive fallacies involved. Accordingly, we see the characterizations provided in the current study as work in progress and, very likely, subject to change.

A closer examination of Propositions 1 to 4 and their proofs reveals that IRs can be categorized into two broad types: *logical-IRs* (denoted by $\overset{l}{\rightsquigarrow}$) and *causal-IRs* (denoted by $\overset{c}{\rightsquigarrow}$). Establishing a logical-IR, $\overset{l}{\rightsquigarrow}$, from a fallacy F_1 to another fallacy F_2 implies that F_1 is a special case of F_2, with every instance of F_1 being an instance of F_2. For example, a closer examination of Proposition 1 and its proof reveals that DE is a special case of FE, with every instance of DE being an instance of FE in disguise. The same understanding holds for Proposition 3 and its proof, indicating that APX is simply a special case of FE, with every instance of APX being an instance of FE in disguise. Hence, using our newly introduced notation: DE $\overset{l}{\rightsquigarrow}$ FE and APX $\overset{l}{\rightsquigarrow}$ FE. Establishing a causal-IR, $\overset{c}{\rightsquigarrow}$, from a fallacy F_1 to another fallacy F_2 implies that the occurrence of F_1 brings about (i.e., causes) the occurrence of F_2. For example, a closer examination of Proposition 2 and its proof reveals that the occurrence of BRN brings about the occurrence of CF, i.e., there is a cause-effect relationship between BRN and CF, with BRN being the cause and CF the effect. The same understanding holds for Proposition 4 and its proof, indicating that the occurrence of AvB brings about the occurrence of CF, i.e., there is a cause-effect relationship between AvB and CF, with AvB being the cause and CF the effect. Hence, using our newly introduced notation: BRN $\overset{c}{\rightsquigarrow}$ CF and AvB $\overset{c}{\rightsquigarrow}$ CF. Drawing further on the analogy between IR and reduction in computational complexity, it is worth noting that there also exist several types of reduction in computational complexity, namely, Karp's reduction, Cook's reduction, truth-table reduction, L-reduction, A-reduction, P-reduction, E-reduction, AP-reduction, PTAS-reduction, etc.

Importantly, logical-IRs and causal-IRs have quite different implications. If $F_1 \overset{l}{\rightsquigarrow} F_2$ holds (implying that F_1 is a special case of F_2 as discussed above), it then follows that a *complete* account of F_2 should also account for F_1, and, in that sense, accounting for F_2 is more demanding[5] than accounting for F_1. For example, since DE is a spacial case of FE (see Proposition 1 and its proof), that is, DE is nothing but FE in disguise, any complete account for FE inevitably should also account for DE, implying that accounting for FE is more demanding than accounting solely for a special case of FE, DE. However, if $F_1 \overset{c}{\rightsquigarrow} F_2$ holds (implying that the occurrence of F_1 brings about F_2), it then follows that an account of F_1 naturally serves as an account of F_2 due to the following rationale: If X causes F_1, and F_1 causes F_2, it then follows that X causes F_2, with F_1 serving as a mediator. In that light, establishing causal-IRs between various cognitive biases/fallacies has an intriguing implication: For any chain of causal-IRs $F_1 \overset{c}{\rightsquigarrow} F_2 \overset{c}{\rightsquigarrow} F_3 \overset{c}{\rightsquigarrow} \ldots \overset{c}{\rightsquigarrow} F_{n-1} \overset{c}{\rightsquigarrow} F_n$, any mechanistic account

[4] We must emphasize that our establishing of the IR between AvB and CF only depends on the broad assumption that the more extreme an event is, the more people overestimate its probability, and holds for any $q(o)$ which satisfies this condition, e.g., $q(o) \propto p(o)|u(o)|\sqrt{\frac{1+|u(o)|\sqrt{s}}{|u(o)|\sqrt{s}}}$ (Nobandegani et al., 2018). Therefore, the assumption $q(o) \propto p(o)|u(o)|$ made in the characterization of AvB is only one choice out of infinitely-many possibilities satisfying the said condition, and hence, is not necessary.

[5] Accounting for F_2 is "more demanding" than for F_1, as a complete account of F_2 would necessarily have to explain a wider range of cases, including all instances of F_1 as a subset.

of F_i naturally serves as an account of $F_{i+1}, F_{i+2}, \ldots, F_n$. For example, since the occurrence of BRN causes the occurrence of CF (see Proposition 2 and its proof), it then follows that any mechanistic account of BRN naturally serves as an account of CF, with BRN serving as a mediator. This understanding has a very intriguing implications for studies of cognitive fallacies in general: Establishing a chain of causal-IRs $F_1 \overset{c}{\rightsquigarrow} F_2 \overset{c}{\rightsquigarrow} F_3 \overset{c}{\rightsquigarrow} \ldots \overset{c}{\rightsquigarrow} F_{n-1} \overset{c}{\rightsquigarrow} F_n$, clearly reveals which of the fallacies $F_1, \cdots, F_n$ is more pivotal or fundamental to account for; the answer is of course the left-most fallacy in the chain, i.e., F_1. This strongly suggests that, directing efforts toward finding a comprehensive, satisfying account of F_1 would be the most rewarding research agenda, because, thanks to the established chain of causal-IRs, we would get a set of comprehensive, satisfying accounts of all $F_2, F_3, \cdots, F_n$ for free! Therefore, identifying IRs could systematize and guide a research agenda, with a huge increase in research efficiency.

Suppose we have established a causal IR between two biases A and B (i.e., $A \overset{c}{\rightsquigarrow} B$). Here is a question worth considering. (Q1) Does a mechanistic account of A also serve as a mechanistic account of B? As we argue above, it does. But it is crucial to note that this is just a theoretical possibility. That is, upon empirical investigations (e.g., using advanced neuroimaging techniques), we may come to realize that the mechanistic underpinnings of B, after all, have nothing to do with that of A. Just because some process model can simulate B does not necessarily imply that that process model is *the* cognitive process responsible for the occurrence of B in the brain. Thus, identifiability remains an issue.

Another question worth considering is the following. (Q2) Assuming we have established $A \overset{c}{\rightsquigarrow} B$, is a participant who commits bias A more likely to commit bias B? The answer to this question is a bit subtle, and is related to our elaboration on (Q1) presented above. If the occurrence of A is indeed what mechanistically drives the occurrence of B (through mechanisms specified in our establishing of the IR between A and B), then the answer to (Q2) is positive. Otherwise, solely based on the fact that we have theoretically established $A \overset{c}{\rightsquigarrow} B$, no decisive answer can be given to (Q2), as there is no *real* mechanistic connection between A and B. Note that just because we have theoretically shown $A \overset{c}{\rightsquigarrow} B$ (i.e., A *can* bring about B, hence a purely theoretical possibility), it does not necessarily imply that A *does* bring about B in reality—the latter claim can be only shown empirically.

Proposition 4, establishing AvB $\rightsquigarrow$ CF, demonstrates an interesting possibility wherein, under a set of auxiliary assumptions (e.g. $\forall i = 1,2,\ |u(o_\wedge)| \gg |u(o_i)|$ in this case), an IR can be established between two fallacies. The idea of establishing IRs under a set of assumptions widens the applicability of the notion of IR, allowing it to link together pairs of cognitive fallacies that would have little connections unless further assumptions are invoked. Drawing again on the analogy between IR and reduction in computational complexity, it is worth noting that in establishing reductions it is common practice to evoke various assumptions/constraints on the characterization of computational problems (e.g. 3-SAT instead of SAT) and/or on the forms of reductions themselves (e.g. *polynomial-time* reductions or *linear-time* reductions). Importantly, these auxiliary assumptions should be empirically confirmed, motivating new and exciting experimental avenues of research. Empirical confirmations of such auxiliary assumptions, empirically justifies the validity of invoking such assumptions. Importantly, empirical disconfirmation of such assumptions, of course, discredit the said established IR, inviting attempts for establishing other IRs (in the hope that they would survive empirical tests), or for invoking other empirically validated assumptions which would save the established IR, motivating new theoretical and empirical work.

In this work, as a proof-of-concept, we establish IRs between several well-documented cognitive biases; see Fig.2(b). Future work should investigate the possibility of establishing IRs between a wider range cognitive biases/fallacies, with the ultimate goal of developing a principled, comprehensive map of cognitive biases/fallacies, broadly resembling what is shown in Fig. 2(a) in the context of computational complexity. As it is conceivable, and in our view very likely, that a single mechanism would act as the common cause of several biases, that mechanism would then serve as a common parent node (in the yet-to-be-developed comprehensive map of biases) having those biases as children. As such, ultimately, the comprehensive map of biases would have (at least) two types of nodes, one to denote biases and one to denote mechanisms.

While many questions remain open, and much work is left to be done in this direction, we hope to have made some progress toward systematically bringing order to the cognitive fallacy zoo. We see our work as a first step in this direction.

Acknowledgments This work is supported by an operating grant to TRS from the Natural Sciences and Engineering Research Council of Canada.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, *21*(4), 503-546. Retrieved from http://www.jstor.org/stable/1907921

Arora, S., & Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Bourgeois-Gironde, S., & Giraud, R. (2009). Framing effects as violations of extensionality. *Theory and Decision*, *67*(4), 385–404.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*(1), 90–98.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 341–350.

Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of computer computations* (pp. 85–103). Springer.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*.

Nobandegani, A. S., da Silva Castanheira, K., Otto, A. R., & Shultz, T. R. (2018). Over-representation of extreme events in decision-making: A rational metacognitive account. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Papadimitriou, C. H. (2003). *Computational Complexity*. John Wiley and Sons Ltd.

Ray, P. (1973). Independence of irrelevant alternatives. *Econometrica: Journal of the Econometric Society*, 987–991.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, *1*(1), 161–176.

Sipser, M. (2006). *Introduction to the Theory of Computation* (Vol. 2). Thomson Course Technology Boston.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981a). *Evidential impact of base rates* (Tech. Rep.). Stanford University, Department of Psychology.

Tversky, A., & Kahneman, D. (1981b). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.