REASONING WITH NEGATIONS: THE AUSTRALIAN PLAN, NEGATION-AS-FAILURE, AND CARD SELECTION TASKS

Francesco Berto and Greg Restall

Arché Philosophical Research Centre
Philosophy Department
The University of St Andrews
Scotland, United Kingdom
[fb96|gr69]@st-andrews.ac.uk*

September 25, 2026

Abstract: Two kinds of formal semantics for negation can help us understand puzzles in the psychology of reasoning: the *Australian Plan semantics*, which generalises the account of negation found in frame semantics for intuitionistic logic, relevant and substructural logics; and *negation-as-failure*, found in logic programming and default logic. We provide a framework where the two kinds of negation coexist, mapping them to the celebrated distinction between System 1 and System 2 thinking. We use the framework to target the phenomenon of *Matching Bias*: specifically, to explain why reasoners fare differently in card selection tasks, depending on whether the target property is phrased positively or negatively.

1. SCENE SETTING

'Logical theorizing helps in understanding and modeling observed behavior in reasoning tasks' (Stenning and van Lambalgen, 2008, p. xiii). We agree and we will put formal logic, and its model-theoretic semantics, to work to investigate how people reason with negations.

There are many ways to understand formal models for a logical system. We could think of them purely instrumentally, as devices we employ to calculate truth

*Both authors contributed equally to this research. ¶ Thanks to audiences at the Arché Metaphysics & Logic Group, the Italian Society for Analytic Philosophy Conference 2023, and at the University of Stirling Philosophy Seminar for feedback on earlier presentations of this work. ¶ Thanks, too, to the referees whose critical comments and questions helped us improve the presentation considerably.

values. Or we could think of them as better or worse representations of reality, where the ideal standard is a model giving us an accurate representation of what is Out There in the Real World. On one reading of that realist view, this is taken to mean that there's only one correct way to model logical phenomena. Opposed to this kind of logical monism is logical pluralism (Beall and Restall, 2000, 2006), according to which there is more than one correct logical consequence relation.

This paper brings together two ways to model the behaviour of negation in logical semantics. But rather than seeing these as representing different, equally good accounts of what follows from what when negation is involved, it uses them to help us understand something of what we *do* when we make inferences with negations. So these different models will be put to work, not to give objective truth conditions for negated sentences, but to shed light on the cognitive processing that might be going on when we reason using negations.

Here is the plan of the paper: in Section 2 we provide the logical background for the work, moving from classical bivalent negation to many-valued semantics; we then present the Australian Plan negation and negation-as-failure. Besides introducing non-specialists to the relevant models, we explain how they can be (and have been) given a natural *cognitive* reading. In Section 3, we present a classical, Wason (1968)-style selection card task and map negation-as-failure to (negation-involving) reasoning in System 1 mode, Australian Plan negation to (ditto) reasoning in System 2 mode. In Section 4, the account is put to work to *explain* how people get the task wrong by sticking with negation-as-failure in System 1 mode, while they get it right by considering information not explicitly given in the task, which is how Australian Plan negation works. In Section 5, we show how the Matching Bias phenomenon, whereby people tend to look for positive information while they are given the corresponding negative, is modeled by the asymmetry between positive and negative information encoded in Australian Plan semantics. Section 6 is a short discussion considering the interaction between conditionals and the processing of negation. We end the paper with Section 7, a methodological coda on how formal modeling can shed explanatory light on empirical results.

2. LOGICAL BACKGROUND

2.1 BOOLEAN TRUTH CONDITIONS

The so-called *Australian Plan* for negation is motivated by considerations in philosophical logic and the semantics for relevant and paraconsistent logics (Berto and Restall, 2019; Meyer and Martin, 1986; Restall, 1999). *Negation-as-failure* is motivated by considerations in logic programming (Clark, 1978; Shepherdson, 1984, 1985). We introduce the two by contrasting them with the traditional truth conditions for negation from classical logic:

- $\neg A$ is true if and only if A is not true.

These can be stated by bringing semantic *values* to the forefront:

- $v(\neg A) = 1$ iff $v(A) \neq 1$

Here, v is an interpretation function assigning truth values to formulas of a language. The Boolean truth conditions assign exactly one of two truth values (here, 1 and 0, but we could equally use T and F, $\circ$ and $\bullet$, or whatnot), in such a way as to ensure that A and $\neg A$ always are assigned different values. One of these values only (here, 1) is taken to stand for truth, and according to any Boolean evaluation v, exactly one of A and $\neg A$ counts as true.

But making semantic values explicit foregrounds another possibility: we might think of there being different values a claim can take, beyond the traditional 1 and 0 for truth and falsity.

2.2 THE AMERICAN PLAN AND THE COGNITIVE READING OF MODELS

A well-known change is to add a third value, say n, for *neither* 1 nor 0, and revise our evaluation clauses to say:

- $v(\neg A) = 1$ iff $v(A) = 0$
- $v(\neg A) = 0$ iff $v(A) = 1$
- $v(\neg A) = n$ iff $v(A) = n$

If having the semantic value n is taken to represent being *neither* true nor false, we have truth value *gaps*. If, on the other hand, it is taken to represent being *both* true and false (so 1 represents being true only, and 0 represents being false only), the three-valued logic allows for truth value *gluts*.

A slightly more involved scheme allows for two intermediate values, one for a gap between truth and falsity, and the other for a glut (Dunn, 1976). A straightforward way to implement a scheme with both gaps and gluts is to understand an evaluation as not supplying a function from sentences to truth-values (whether there are two, three or more of them), but a relation to the original values 0 and 1. We have:

- $\neg A \rho 1$ iff $A \rho 0$
- $\neg A \rho 0$ iff $A \rho 1$

This is to be understood as saying that the negation $\neg A$ is true (it is related by ρ to 1) if and only if A is false (it is related by ρ to 0) and $\neg A$ is false if and only if A is true (Priest, 2008). A is both true and false if $A\rho 1$ and $A\rho 0$ holds and it is neither true nor false if $A\rho 1$ and $A\rho 0$ both fail.

This way of modeling negation has become known as the *American Plan* (Routley, 1984). One popular interpretation of the four values takes them to be *epistemic* or, if one likes, *cognitive*. This means that we don't take a sentence that's related to both 1 and 0 as telling us about a contradiction that's out there, realized in the world; or a sentence that's related to neither 1 nor 0 as telling us about some underdeterminacy of reality. Rather, we take such sentences as marking inconsistent and/or incomplete *data* or *information*. Given some database of evidence available to an agent, represented in a mind, stored in some artificial system, or whatnot – especially evidence which may come from disparate sources –, it is natural to classify declarative sentences in four ways. The database may declare a sentence A true (only); it may declare A false (only). But the database may also stay silent on the issue of whether A or not. Or, the database may contain conflicting information, and declare A both to be true and to be false (Belnap, 1977a,b; Hazen and Pelletier, 2019). We won't proceed with three- or four-valued logics from now on; but we will retain an insight from this work: that some logical formalisms can be used to represent, not (only) the workings of truth and falsity as such, but what is supported by a state of information.

2.3 THE AUSTRALIAN PLAN

We can state the Boolean truth conditions also by bringing *scenarios* to the forefront:

- $s \Vdash \neg A$ iff $s \nVdash A$

Read '$s \Vdash A$', in general fashion, as 'scenario s supports A'. A scenario is formally represented just as a point in a model, at which formulas are evaluated, but soon we will say more on what scenarios could *be*. When we evaluate statements with respect to each scenario, the traditional Boolean conditions for negation are straightforward: a negation holds at a scenario if and only if the thing negated does not hold. But making the scenario parameter explicit foregrounds another possibility: we might think of there being different scenarios where formulas can be evaluated, allowing our models to contain more than one.

Now the *Australian Plan* semantics for negation analyses the truth of $\neg A$ at a scenario s in terms of whether A holds at *other*, related scenarios (Berto, 2015; Berto and Restall, 2019). The idea has a long and venerable heritage. In Beth or Kripke semantics for intuitionistic logic (Beth, 1956; Kripke, 1965), scenarios are taken as *constructions* of some kind, e.g., proofs produced by Brouwer's 'creating subject' or idealized mathematician. A construction c supports $\neg A$ if and only if there is no construction c' extending c, which supports A. So, if c does not yet have the power to verify A, but some extension c' of c does manage the task, then c neither supports A nor $\neg A$.

Intuitionistic logic is but one example of a system that fits the Australian Plan semantics framework. The terminology was introduced in the semantics for *relevant logics*, with an idea due to Richard Sylvan and Valerie Plumwood (then writing as Richard and Valerie Routley) (Routley and Routley, 1972), according to which a scenario s (they called them 'set-ups') supports $\neg A$ if and only if its *star*-mate s^* does not support A. Provided that s and s^* differ, we could have incomplete set-ups, supporting neither A nor $\neg A$, and inconsistent ones, supporting both A and $\neg A$. (What's the star-mate of a scenario? Well, we'll soon have more to say on what scenarios could be; but an intuitive idea of the relation between s and s^* is that the star-mate of a scenario reverses the relevant scenario in-and-out (Dunn, 1994): either gives inconsistent information where the other gives incomplete one, so if either supports both A and $\neg A$, the other will support neither.)

We do not need to go into the detail of the Routley star here (Dunn, 1994; Restall, 1999; Routley, 1984), except to note that it, as well as the semantics for intuitionistic logic, can both be generalised as the Australian Plan semantics for negation. The key notion at work is a binary *compatibility* relation C on scenarios:

- $s \Vdash \neg A$ if and only if for each t where sCt, $t \nVdash A$

A scenario s supports $\neg A$ iff every scenario t *compatible* with s does not support A. Or, in other words:

- $s \Vdash \neg A$ if and only if for any t where $t \Vdash A$, not sCt

A scenario s supports $\neg A$ iff any scenario that supports A is *incompatible* with s.

The core idea behind the Australian Plan is obvious: we utter negations to express incompatibilities, i.e., to rule out that something is the case (Kinkaid, 2020). The way we experience the world tells us that it's full of incompatibles: given the physics of light wavelength reflection, *being red all over* rules out *being blue all over* at the same time. The way numbers are, *Being odd* rules out *being even*. Given the nature of spacetime, *being here* is incompatible with *being there* at the same time, for suitable 'here' and 'there'. So to say that, e.g., the chair is *not* red is to rule out that it is red. What do we know when we know, of a chair, that it's not red? Presumably, that it has some feature or other such as a different colour, *incompatible* with being red. To say that Fred is *not* in the kitchen is to rule out that Fred is in the kitchen. What do we know when we know, of Fred, that he's not in the kitchen? Presumably, that he has some feature, such as being somewhere or other, *incompatible* with being in the kitchen. It stands to reason that humans and other sentient beings might naturally develop systems of representation with a device designed to record and communicate to each other such incompatibilities. 'Not' is what does the trick for us. For think about how dysfunctional a language would be, if it lacked an exclusion-expressing device:

> *Me*: 'Fred is in the kitchen.' (Sets off for kitchen.)
>
> *You*: 'Wait! Fred is in the garden.'
>
> *Me*: 'I see. But he is in the kitchen, so I'll go there.' (Sets off.)
>
> *You*: 'You lack understanding. The kitchen is Fred-free.'
>
> *Me*: 'Is it really? But Fred's in it, and that's the important thing. (Leaves for kitchen.).
>
> Your problem is to get me to appreciate that your claims are incompatible with mine. (Price, 1990, p. 224)

Now this makes sense given a wide range of interpretations of what counts as a *scenario*. Say, we restrict our attention to taking scenarios as *classically possible worlds* (i.e., in the interpretation of standard modal logic, maximally consistent ways things could be or have been). Then we recover the original Boolean semantics, because any two different worlds are pairwise incompatible.

But the Australian Plan clause makes sense for richer collections of scenarios. One can take these as restricted parts of the world, like the *situations* of Barwise and Perry's situation semantics (Barwise and Perry, 1983); or as constructions, as in constructivist and, in particular, intuitionistic logic; or as information states, as in (certain interpretations of) relevant logic; or as bodies of evidence, etc. If a scenario s supports $\neg A$ and another scenario t supports A, then because negation records incompatibilities, s and t are incompatible: one rules A in and the other rules A out. On the other hand, if s does *not* rule A out, then if our family of scenarios is rich enough, it is plausible to take there to be *some* scenario, compatible with s, that supports A. Otherwise, it seems that the original scenario did indeed rule A out after all.

One salient feature of the Plan is its *modal* character, in the sense of modal logic (and as recorded in the modal flavour of 'incompat*ible*', i.e., 'such that it *can't* be together with'): evaluating a negation at a scenario s can involve considering *other* scenarios, different from s, (in)compatible with the original one. Asked whether the chair is red, we may say: I know it's blue, and that's incompatible with its being red. Asked whether Fred is in the kitchen, we may reply: Well I see him in the garden, so there's no way he's in the kitchen.

In this way, the semantics of negation is *intensional* in character, and is not best characterised as 'truth-functional' or *extensional* even when we expand the range of possible truth values beyond the traditional 'true' and 'false'. The intensional approach to negation parallels the intensional treatment of the conditional, which has a long history, going back, in modern times, at least to Lewis' calculus of strict implication (Lewis, 1914). Here, however, the core issue is not whether negation might be objectively extensional (here, we take no issue with the traditional view that $\neg A$ is true if and only if A is not true), but how we process conditionals in realistic epistemic scenarios: a limited scenario that does not support A might well not support $\neg A$ either, and our reasoning in such a scenario is properly intensional.

If one takes thinking about alternative scenarios to have some *cognitive cost*, the task of assessing an Australian Plan negation might be cognitively demanding, in that it may involve considering a range of alternatives: we may conclude that $\neg A$ is supported at our s, only after considering that various scenarios which we find compatible with s are such that A fails at them; or, equivalently, only after considering that other scenarios which we find incompatible with s are such that A holds at them.

Now there is an amount of empirical evidence from the psychology of reasoning, showing that the consideration of a plurality of scenarios is indeed a cognitive cost. Such evidence, reported e.g. in (Johnson-Laird and Byrne, 1991), is used by mental model theorists like Philip Johnson-Laird and Ruth Byrne to explain errors of reasoning with conditionals, like the ones we will meet later when we come to card selection tasks. As Ruth Byrne has it, 'Multiple possibilities tend to exceed working-memory capacity'; and 'the greater number of possibilities increases the difficulty of the inference' (Byrne, 2005, pp. 21 and 25). This feature will pay a crucial role when we reflect on the difficulties of reasoning with negations.

2.4 NEGATION AS FAILURE

When it comes to epistemic, evidential or informational understandings of the semantics of negation, there is another important tradition: that of *logic programming* in knowledge representation and database theory. A question tackled in this tradition is: when we have a database filled with individual pieces of information (concerning, for example, which student has what student number, what modules they have been enrolled in, and what grades they have in each module), when does such a database support a negative judgement?

One influential proposal is *negation-as-failure* (Clark, 1978; Reiter, 1978; Shepherdson, 1984, 1985). Back to the Boolean clause for negation:

- $s \Vdash \neg A$ iff $s \nVdash A$

On a cognitive-informational reading of the semantics, this would seem more straightforward to process than the Australian Plan clause, in that to assess the status of $\neg A$ at a scenario s we only consider the information supported by s itself. Suppose s is given by a database. If we take it to contain all of the (salient) information about the items in question, this makes perfect sense. If the information that the student 123456789 is enrolled in PY2010 in 2024 *fails* to show up in s, then the right-to-left direction of the biconditional above allows us to conclude that, according to s, it's not the case that student 123456789 is enrolled in PY2010 in 2024.

The same goes for evidential scenarios and phenomenally salient properties. If we can see a shape before us, and it fails to appear to be square, then that shape appears to *not* be square. In the context of logic programming, one usually specifies that the negation-as-failure of A is supported by a database when the attempt to extract A from the database fails in a number of steps (which may be rather small). When one sees a database as what's represented by a scenario, the crucial idea is that we don't look beyond it, i.e., we stick to whatever, if little, information is supported by the scenario. Stenning and van Lambalgen (2008) rightly call this 'closed-world reasoning'; see also the 'closed world assumption' known from logic programming: Reiter (1978); Shepherdson (1984).

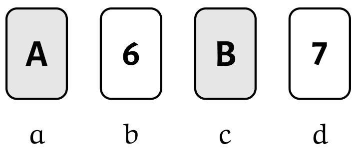
So far, so good—but, the negation-as-failure clause might misfire when the scenario is incomplete. And one may claim that *all* of our epistemic scenarios (bodies of evidence, available information, etc.) in ordinary life are by default incomplete, as they do not contain enough information to decide every issue. Negation-as-failure could then be seen, cognitively speaking, as a rough-and-ready evaluation strategy that requires none of the process of looking at alternative scenarios.

(In fact, one way of understanding negation-as-failure is not as a rival to the Australian Plan but as an *instance* of it, in which the search for alternative scenarios is cut short by considering only the scenario of evaluation.) Negation-as-failure pays for this simplicity by over-generating, producing false positives (cases where the negation is regarded as true when it should be regarded as untrue) when applied to incomplete scenarios, where a more careful evaluation strategy would not.

This does not mean that negation-as-failure is useless. On the contrary, it may be (and, has been: Stenning and van Lambalgen (2008) again) taken as modelling how rational agents with bounded time and cognitive resources can efficiently extract negative information from limited evidence. When we read the train timetable and find it lacks notice of trains leaving for Edinburgh between 10 and 11, it would be crazy to refrain from concluding that it's *not* the case that some train leaves for Edinburgh between 10 and 11 until we have complete information on the status of the whole of reality. Oftentimes, we have no time or resources to even consider any alternative compatible situations, let alone, to systematically enumerate them all.

3. TAKING TWO PERSPECTIVES

So from now on, take our scenarios as evidential situations, representing the information available to a reasoner. Here's a situation where we see the following four cards, each of which has a letter on one side and a numeral on the other. The cards are labelled for ease of reference:



Two cards show their letter side, card a shows an A, and card c shows a B. The other two cards show their number side. Card b shows a 6, while card d shows a 7. We cannot see the numerals on the other sides of cards a and c or the letters on the other sides of cards b and d. We will consider evidential situations like these as simple, concrete cases, in which our reasoning operates. This kind of situation is familiar from card selection tasks à la Wason (1968), which Stenning and van Lambalgen call 'the mother of all reasoning tasks' (Stenning and van Lambalgen, 2008, p. 44).

We can ask ourselves whether, in evidential situation s, a judgement A is *correct*. But this would be to miss the point of restricting our attention to the evidence available to the agent. After all, in this scenario there is some face value that card a has, it is just not known by the agent. Once we have given the *basic* information available to the agent, we can ask, what further, logically *complex* information is thereby available? This depends on what the agent can *do*: on their cognitive resources and capacities.

Now there's a celebrated distinction between two modalities of thought, traced back to the work of authors like Kahneman and Tversky, Evans and Over (Kahneman, 2011; Solaki et al., 2019; Raoelison et al., 2021; Evans and Over, 2004): thinking in System 1 mode, also popularized as 'fast thinking', and thinking in System 2 mode, also popularized as 'slow thinking'. Quick ways of drawing the distinction from the literature have it that System 1 does not apply formal rules of logic or probability, works by associative memory and can import biases and intuitions; whereas System 2 is at work in logical deductions or probabilistic estimates, follows precise rules, etc. But perhaps the two most salient differences between the two are (a) that System 1 is automatic, whereas System 2 can be activated deliberately; and (b) that System 1 is cognitively effortless, whereas using System 2 requires paying cognitive costs.

In addition, the Kahneman-Tversky narrative has it that we should not take System 1 as the rationally bad guy, and System 2 as the rationally good guy: we could not navigate life in System 2 mode all the time, as that's cognitively unfeasible. System 1 gets things right, usually – it is just prone to errors, of the kind exemplified, e.g., by reasoning fallacies, in specific circumstances. And System 2 can be 'lazy', not stepping in to fix the mistaken judgments of System 1 when it should.

So let's define two evaluation relations on epistemic scenarios, which we take as corresponding, respectively, to thinking in System 1 and System 2 mode, $\Vdash_1$ and $\Vdash_2$:

- '$s \Vdash_1 A$' says: an agent has the capacity to derive A in epistemic scenario s, using System 1, or: s supports A for an agent via System 1 thinking.
- '$s \Vdash_2 A$' says: an agent has the capacity to derive A in epistemic scenario s, using System 2, or: s supports A for an agent via System 2 thinking.

We can assume that System 1 and System 2 operate on the same set of *basic* information in each scenario. This is a safe idealization: it will not make any difference in what follows, since our focus is the behaviour of negation. We have, for atomic sentences p:

- $s \Vdash_1 p$ if and only if $s \Vdash p$
- $s \Vdash_2 p$ if and only if $s \Vdash p$

Where $\Vdash$ is the given notion of basic evidence in a scenario. The key difference between the two support relations is their treatment of negation: (we suppose you had already guessed that) we propose System 1 support is given by negation-as-failure:

- $s \Vdash_1 \neg A$ if and only if $s \nVdash_1 A$

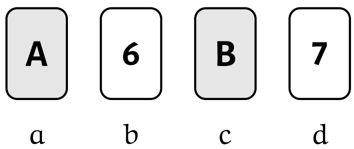
And System 2 support uses the notion of compatibility between evidential situations, in Australian Plan fashion:

- $s \Vdash_2 \neg A$ if and only if for each s', where sCs', $s' \nVdash_2 A$.

This more complicated support relation requires quantification over evidential scenarios, and the details of how the compatibility relation is to be defined will depend on how those scenarios are specified. To show how this might be done, we will consider a concrete example in the next section. Before we get there, however, it is worth reiterating what this formalism represents, and what it does not represent. First, despite there being two different semantic evaluation clauses for negation, the formalism is not encoding two different *concepts* of negation. The difference between the two systems is represented by the two different support relations, representing the capacities an agent has using System 1 and System 2, respectively. It is not that System 1 judgements are about one kind of negation, and System 2 judgements are about another. Second, we make no empirical claim about exactly *how* System 1 and System 2 processing is cognitively encoded. It may well be that System 1 negation processing is a cut-down version of System 2 (in which the reasoner does not consider a range of alternative scenarios), and so, the situation is akin to Egré's 'parametric monism' (Égré, 2021), in which we move from one evaluation process to another by varying a parameter.

4. CARD SELECTION TASKS

Back to the card selection scenario discussed above:



Let's say this is a scenario s with objects $\{a, b, c, d\}$ and properties assigned as follows:

s	A	B	6	7
a	1	—	—	—
b	—	—	1	—
c	—	1	—	—
d	—	—	—	1

This table is one way to represent the following four basic evidential judgements, the raw materials on which Systems 1 and 2 operate:

$$s \Vdash Aa \qquad s \Vdash 6b \qquad s \Vdash Bc \qquad s \Vdash 7d$$

That's just what we *see*, looking at the four cards. System 1 operates on this situation in the expected way. For example, concerning the first card, we have:

$$s \nVdash_1 \neg Aa \qquad s \Vdash_1 \neg Ba \qquad s \Vdash_1 \neg 6a \qquad s \Vdash_1 \neg 7a$$

The first card does not have a B (indeed, it has a A), and as far as System 1 is concerned, it isn't a 6 and it isn't a 7, either. Of course, we know that it *might* be a 6 and it *might* be a 7 for all we know. But to say 'might' is to go modal, i.e., to use System 2: that judgement appeals to alternative scenarios. We can consider at least two different possibilities, where we have seen both sides of the first card, revealing either a 6 or a 7. Call these scenarios s_{a6} and s_{a7}, respectively:

s_{a6}	A	B	6	7
a	1	–	1	–
b	–	–	1	–
c	–	1	–	–
d	–	–	–	1

s_{a7}	A	B	6	7
a	1	–	–	1
b	–	–	1	–
c	–	1	–	–
d	–	–	–	1

We are now considering three different scenarios. Both new scenarios s_{a6} and s_{a7} are compatible with the original s, since they just fill in some information that s left out. But they are *incompatible* with each other. This is part of our background knowledge of how these cards are set up: no card can be both a 6 and a 7.

In our simple card-based setup, we can say that two scenarios are compatible if and only if, whenever one scenario takes a card to have one letter [number], the other scenario does not take that card to have a *different* letter [number]. (Notice that on this view, if we have a confused scenario s that takes a card to have two different numbers, this can be compatible with another scenario t that does not ascribe any number to that card. This seems right to us: s does not clash with this t, even though it does clash with *itself*.) So, we have sCs_{a6} and sCs_{a7} (and, of course, sCs) but we do *not* have $s_{a6}Cs_{a7}$. With these compatibility facts in hand we get:

$$s \nVdash_2 \neg Aa \qquad s \Vdash_2 \neg Ba \qquad s \nVdash_2 \neg 6a \qquad s \nVdash_2 \neg 7a$$

Since s is compatible with itself and $s \Vdash_2 Aa$, we have $s \nVdash_2 \neg Aa$. But *any* scenario t where $t \Vdash Ba$ would be incompatible with s (since s already takes a to be A), so we do have $s \Vdash_2 \neg Ba$, as we would hope. So far, $\Vdash_1$ and $\Vdash_2$ agree. The extra power of System 2 reasoning comes to the fore concerning the judgements $\neg 6a$ and $\neg 7a$. Since $s_{6a} \Vdash_2 6a$ and sCs_{6a}, we have $s \nVdash_2 \neg 6a$, and similarly, since $s_{7a} \Vdash_2 7a$ and sCs_{7a}, we have $s \nVdash_2 \neg 7a$. By considering alternate scenarios, System 2 is able to find counterexamples to both negative claims.

5. CARD SELECTION AND THE MATCHING BIAS

The *Matching Bias* is a well-documented phenomenon in the psychology of reasoning, whereby reasoners tend to look for positive information while what they are explicitly given is the corresponding negative information. It started with variations on the Wason (1968) selection task: people are asked to test conditional claims like 'If there's an A on one side of the card, then there's a 7 on the other side' on four cards which have a letter on one side and a single digit on the other. Up to 90% of people tend to select cards with A and 7. But they should choose cards with an A and *without* a 7: a search to falsify the conditional is a search for a situation that makes the antecedent true and the consequent false.

But Evans (1972, 1998) and others found that, if you put a negation in the consequent, people's performance drastically increases: when the instruction is to test 'If there's an *A* on one side of the card then there's not a 7 on the other side', almost everyone gets the task right. In this case the Matching Bias, i.e., the fact that people look for a card with a 7 when they are given the information 'not-7' in the consequent, helps them: 'adding a negation has made the task much easier' (Evans and Over, 2004, p. 75).

One explanation for the Matching Bias, advanced by Oaksford and Chater (1994) and others, is that people have an *expected positive information gain*. The hypothesis is that there is an asymmetry between positive and negative information: negative claims are more difficult to process than positive ones, because they refer to relatively large contrast classes and so provide low information (and the lower, the larger the contrast class is). This, as well as the Matching Bias, has received some empirical corroboration (Yama, 2001).

This asymmetry hypothesis (and its empirical corroboration) fits more naturally with the Australian Plan semantics for negation than with the American Plan, *if* these are understood in terms of cognitive processing. In the American Plan, positive and negative information are perfectly symmetric: negation is just a flip-flop between truth and falsity, treated even-handedly.

But the asymmetry hypothesis maps neatly to negation on the Australian Plan, plus the idea that dealing with negative information in Australian Plan fashion requires System 2 effort. One core idea of the Australian Plan is that we use negations to express incompatibilities or exclusions. So 'not-*A*' is understood as 'Something is the case, which rules out that *A* is the case' or, read more epistemically, as we should when we want to track cognitive phenomena in the psychology of reasoning, 'Some accessible/available information rules out that *A* is the case'. In particular, when we process negations in System 2 mode by the Australian Plan rule, we may often have to mentally check a plurality of (in)compatible scenarios.

But then, the Matching Bias effect should be reduced to the extent that the contrast class (we'd say: the class of incompatible scenarios) one needs to consider gets smaller in size. When we process negations in System 1 mode, we just look at the information that's directly given to us in a scenario and apply negation-as-failure: we take not-*A* as true (resp. false) when *A* is absent (resp., is present) in the relevant database. Whereas when we process them in System 2 mode, we put in the cognitive effort required to go beyond the info listed in the scenario by considering alternatives.

As soon as we add a negation in the consequent of the conditional to be tested in the standard Selection Task, 'If there is an *A* on on one side, then there is not a 7 on the other', then the task of searching for a counterexample becomes easy. That's because the contrast class (any card with a number different from 7 on it) is now explicitly given. The original Selection Task asked us to reason thus: 'Search for a card with *A* on one side, and *some number incompatible with* (having a) 7 on the other'. That is a relatively large contrast class, tracked via the notion of incompatibility: one has to think through a number of different situations incompatible with having a 7: having a 1, a 2, a 3, … etc., to extract the information that one is to select non-7 cards. When a negation is added in the consequent, negation-as-failure in System 1 mode is sufficient to get the right answer: the inference is from the presence of a 7 in the basic information (the card with a 7 is there to see) to the falsity of non-7.

6. CONDITIONALS, PROBABILITIES, AND NEGATION

We now focus on issues concerning negation and the processing of conditional sentences in general, beyond card selection tasks. While this is such a broad topic that it would deserve (at least) one separate paper, there is a sub-topic which deserves further discussion here.

Experimental results summarized and discussed in Over and Evans (2024) concern the fact that subjects comply with the so-called Ramsey Test in their 'suppositional' assessment of conditionals: they generally take the probability of a conditional, $P(A \to C)$, as the probability of the consequent conditional on the antecedent, $P(C \mid A)$, which one can read as: on the supposition of the antecedent. Subjects do it, in particular, with all combinations of the presence and absence of negations in the antecedent and consequent. So people tend to treat conditionals with opposite consequents, $A \to C$ and $A \to \neg C$, as complementary, as if they are contradictories (see Handley et al., 2006). This goes hand in hand with the phenomenon, acknowledged and discussed e.g. in Chapter 9 of Evans and Over (2004), whereby people tend to naturally switch between external and internal negation (in the consequent), treating $\neg(A \to C)$ as interchangeable with $A \to \neg C$.

Now a helpful referee asks: how do negation-as-failure and Australian Plan negation relate to these results? In particular: what about $P(\neg C \mid A)$ when 'If A, then not-C' is read as saying that, given a construction supporting A, there is no construction extending it and supporting C? This is a welcome question, for the aforementioned experimental findings are naturally accommodated by our mapping negation-as-failure to System 1 and Australian Plan negation to System 2.

Negation-as-failure complies with complementarity: that the negation-as-failure of C, $\neg C$, holds in scenario s means that the information that C is not supported in the relevant, task-closed scenario s. When assessing 'If A, then not-C', by the Ramsey Test the subject will assign to it a probability that's $P(\neg C \mid A)$. When scenario s is looked at on the supposition of A, this will be treated as closed for the task. Since $\neg C$ behaves like the classical complement of C, we have $P(\neg C \mid A) = 1 - P(C \mid A)$, and so the pair of $A \to C$ and $A \to \neg C$ will be taken as complementary.

As for the Australian Plan negation, in its general setting it says little on the nature of *scenarios*: it's a very general, structural framework that can accommodate different logical accounts of negation – as soon as one says more on what one wants a scenario to be. If one wants scenarios to be taken as incomplete-extendible constructions, as per our helpful referee's question, there can be such A-scenarios (with nonzero probability assigned) that support neither C nor $\neg C$. We can then have that $P(C \mid A) + P(\neg C \mid A) < 1$, and so, by the Ramsey Test, $P(A \rightarrow C) + P(A \rightarrow \neg C) < 1$, and so complementarity fails in general, although the Ramsey Test can still hold separately for each conditional.

Now our mapping of negation-as-failure to System 1 and of Australian Plan negation to System 2 allows for the following story: As remarked in Evans and Over's dual-process view of our assessment of conditionals, suppositional processing via Ramsey Test oftentimes relies on 'preconscious heuristic (System 1) processes' (see Evans and Over, 2004, 76), rather than analytic System 2 reasoning. So when subjects, as they often do, assess in System 1 mode a conditional with a negation in the consequent, one can conjecture, subjects aren't performing the cognitively expensive universal quantification over a plurality of compatible scenarios Australian Plan semantics points at. They are doing, rather, negation-as-failure-style absence-checking and running the Ramsey Test suppositional procedure with it.

7. THE UPSHOT

This is just the beginning of an exploration: we have been tracking but a tiny slice of the huge amount of empirical literature on selection tasks and the Matching Bias for negation from cognitive psychology. One important issue we have not dealt with, for instance, is *deontic* selection tasks. It is well-known that, in general, people fare better with selection tasks when deontic conditionals are involved. With deontics like (as per a helpful referee's example) 'If customers in shops are sold alcohol, then they must not be under 18 years old', participants have no problem choosing the 'sold alcohol' card and the '17 years old' card.

Now of course, because the referee's example has a negated consequent, the matching bias can already be part of an explanation of why the task is easier. But, because there's a 'must' in the consequent, the relevant logical form is not just 'If A, then not-C', but 'If A, then M: not-C', where 'M: ...' is 'It must be the case that: ...'. We agree with Stenning and van Lambalgen (2008), Chapter 3, that the specific logical form of deontics should be taken into account in assessing what goes on here. If one (a) takes deontic operators as modals, i.e., as quantifiers over scenarios, as is standard in model-theoretic logical semantics, and (b) takes their psychological role in conditionals as that of instructions to look for rule violators, as Evans and Over (2004) seem to us to suggest (see p. 82–83), then deontic operators may make the relevant violation condition explicit. Deontic tasks with negations embedded in the consequent do not challenge our account; rather, they combine negated-consequent facilitation with pragmatic and normative violation-detection mechanisms primed by deontic modals.

We have focused on the *explanatory* value of some formal logical modeling: if we take on board the System 1/System 2 distinction, and we map it to our two proposed logical formalisms for negation, our model will explain something about people's reasoning behaviour with negations. We even refrained from extracting testable *predictions* from the model; although we think this could be done, it is not us 'armchair' logicians who are positioned to do it.

So we would like our reader to see this little paper as a proof-of-concept: a little (formal) logic goes a long way. That's because it gives us idealized models of information processing systems—'idealized', as in: '*simplified*', not perforce as in: 'having *normative* value', i.e., representing what we ought rationally to do by representing what an ideal rational reasoner does. That doesn't rule out that the models may also have some normative value. But logicians needn't take a stance on whether that's the case: they won't wait for theorists engaged in the so-called rationality debate (Stein, 1997) to reach an agreement on the Ultimate Nature of Rationality, before they start building their models.

As Daniel Greco (2023) has convincingly argued, the opposition between descriptive and normative for idealized models in formal epistemology is sometimes a false dichotomy. It stems, for instance, from the idea that logic is normative for reasoning whereas the psychology of reasoning is merely descriptive of thinking processes. But on the one hand, it is highly controversial whether logic *is* normative (Steinberger, 2022; Russell, 2017); on the other hand, such processes may sometimes be successfully represented by models that also have normative value, under the not altogether implausible assumption that what we usually do is also often what we ought to do, given our limited resources of time, memory, and processing powers.

As Evans and Over have it:

> In examining psychological evidence, our difficulty will not be lack of evidence: indeed there is a wealth of studies. The problem is that psychological experiments are a lot more complex to interpret than might at first be thought and introduce numerous unexpected and often perplexing findings. (Evans and Over, 2004, p. 34)

What we sometimes need is *explanations* for the findings. Formal logical models, idealized in the relevant sense, are successful to the extent that they filter out most real-world psychological complexities, and give us a simple representation unifying a variety of observed phenomena: a core feature of a theory having explanatory value, on one traditional way of understanding explanations (Kitcher, 1976).

So we have offered our two-level analysis of negation-processing as a way to use simple tools from contemporary logic to understand different ways negative information might be processed, and to see how this fits with the empirical data available in the psychology of reasoning. This is something that the theoreticians can offer to the experimental psychologists: we conjecture that some of these modeling tools may have explanatory power. We think the kind of explanation we have given here shows that logicians' work (even work on seemingly exotic formal systems) can find application not just in the world of semantics, as an account of the real-world truth conditions of logical concepts, but as a description of a process that could be followed by a reasoning agent. You do not need to think that two-valued truth tables are all there is to say about processing negations. We hope that considerations of this kind may generate more traffic between the formal theoreticians and the cognitive psychologists.

Further, the structure of this analysis points to a number of related issues to explore. On our view the System 2 process for evaluating negations involves a universal quantification. These are, themselves, notoriously subtle, with their own scope/domain restriction phenomena. We expect to find some kind of connection between negation processing and contextually determined quantifier scope restrictions, when agents attempt to consider different circumstances compatible with their current evidence.

Another related issue is the cognitive processing of other notions that also plausibly involve comparison of alternative scenarios, and how these notions are processed in System 1 and System 2. The epistemic 'might' modality is a natural candidate for future work.

Finally, we think that there are also lessons here for logicians like us. We have learned something when we took a step back and thought of negation on the Australian Plan and of negation-as-failure, not just as accounts of when a negation is true, but as ways of accounting for how agents assess negative claims relative to their own information states. This shift of perspective is fruitful—as we hope to have shown—but it is not one that comes easily. We are trained to read truth conditions as real-world descriptions of what is true, and so, when confronted with the difference between the Australian Plan and negation-as-failure, for example, we are trained to think of one of them as correct and the other as incorrect. If they are understood, instead, as representing different things we might do when reasoning with negations, we have the means to see how many different structures find their place as representations of distinct kinds of cognitive processing. We expect that if logicians learn to see their work in this way (in addition to all the other ways logic has been applied, not necessarily as a replacement), then there will be many more insights ahead.

DECLARATIONS

ETHICS APPROVAL

Ethics approval was not required for this research.

FUNDING

This research was supported by the Leverhulme Trust.

CONFLICT OF INTEREST/COMPETING INTEREST

The authors have no conflicts of interests to declare.

DATA AVAILABILITY

This research did not require the collection of data.

REFERENCES

J. Barwise and J. Perry. *Situations and Attitudes*. CSLI Publications, Stanford, CA, 1983.

JC Beall and Greg Restall. Logical pluralism. *Australasian Journal of Philosophy*, 78:475–493, 2000.

JC Beall and Greg Restall. *Logical Pluralism*. Oxford University Press, Oxford, 2006.

Nuel D. Belnap. A useful four-valued logic. In J. Michael Dunn and George Epstein, editors, *Modern Uses of Multiple-Valued Logics*, pages 8–37. Reidel, Dordrecht, 1977a.

Nuel D. Belnap. How a computer should think. In G. Ryle, editor, *Contemporary Aspects of Philosophy*, pages 30–55. Oriel Press, 1977b.

Francesco Berto. A modality called "negation". *Mind*, 124(495):761–793, 2015.

Francesco Berto and Greg Restall. Negation on the Australian plan. *Journal of Philosophical Logic*, 48(6):1119–1144, 2019.

Evert W. Beth. Semantic construction of intuitionistic logic. *Koninklijke Nederlandse Akademie van Wentenschappen, Proceedings of the Section of Sciences*, 19:357–388, 1956.

R. Byrne. *The Rational Imagination. How People Create Alternatives to Reality*. MIT Press, Cambridge, MA, 2005.

Keith L. Clark. Negation as failure. In Hervé Gallaire and Jack Minker, editors, *Logic and Data Bases*, pages 293–322. Plenum Press, 1978.

J. Michael Dunn. Intuitive semantics for first-degree entailments and "coupled trees". *Philosophical Studies*, 29(3):149–168, 1976.

J. Michael Dunn. Star and perp: Two treatments of negation. In James E. Tomberlin, editor, *Philosophical Perspectives*, volume 7, pages 331–357. Ridgeview Publishing Company, Atascadero, California, 1994.

P. Égré. Logiques non-classiques et pluralisme logique. In Francesca Poggiolesi and Pierre Wagner, editors, *Précis de philosophie de la logique et des mathématiques. Vol. 1 Précis de philosophie de la logique*, pages 377–421. Editions de la Sorbonne, 2021.

J. Evans. Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology*, 24:193–99, 1972.

J. Evans. Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Resasoning*, 4:45–82, 1998.

J. Evans and D. Over. *If*. Oxford University Press, Oxford, 2004.

D. Greco. *Idealization in Epistemology. A Modest Modeling Approach*. Oxford University Press, Oxford, 2023.

S. Handley, J. Evans, and V. Thompson. The negated conditional: A litmus test for the suppositional conditional? *Journal of Experimental Psychology*, 32:559–569, 2006.

Allen P. Hazen and Francis Jeffry Pelletier. K3, Ł3, LP, RM3, A3, FDE, M: How to make many-valued logics work for you. In Hitoshi Omori and Heinrich Wansing, editors, *New Essays on Belnap–Dunn Logic*, pages 155–190. Springer International Publishing, 2019. ISBN 978-3-030-31136-0.

P. Johnson-Laird and R. Byrne. *Deduction*. Erlbaum, Hillsdale, NJ, 1991.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. ISBN 0374275637.

J. Kinkaid. What would a phenomenology of logic look like? *Mind*, 129:1009–1031, 2020.

P. Kitcher. Explanation, conjunction, and unification. *Journal of Philosophy*, 73:207–212, 1976.

S. A. Kripke. Semantical analysis of intuitionistic logic. In J. Crossley and M. A. E. Dummett, editors, *Formal Systems and Recursive Functions*, pages 92–130. North-Holland Publishing, Amsterdam, 1965.

C. I. Lewis. The calculus of strict implication. *Mind*, 23(90):240–247, 1914.

Robert K. Meyer and Errol P. Martin. Logic on the Australian plan. *Journal of Philosophical Logic*, 15(3):305–332, 1986.

Mike Oaksford and Nick Chater. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4):608–631, 1994.

D. Over and J. Evans. *Human Reasoning*. Cambridge University Press, 2024.

Huw Price. Why 'not'? *Mind*, 99:221–38, 1990.

Graham Priest. *An Introduction to Non-Classical Logic: from if to is*. Cambridge University Press, Cambridge, 2008. ISBN 0521670268.

Matthieu Raoelison, Esther Boissin, Grégoire Borst, and Wim De Neys. From slow to fast logic: the development of logical intuitions. *Thinking & Reasoning*, 27(4):599–622, 2021.

Raymond Reiter. On closed world data bases. In Hervé Gallaire and Jack Minker, editors, *Logic and Data Bases*, pages 55–76. Plenum Press, 1978.

Greg Restall. Negation in Relevant Logics: How I Stopped Worrying and Learned to Love the Routley Star. In Dov Gabbay and Heinrich Wansing, editors, *What is Negation?*, volume 13 of *Applied Logic Series*, pages 53–76. Kluwer Academic Publishers, 1999.

Richard Routley. The American plan completed: Alternative classical-style semantics, without stars, for relevant and paraconsistent logics. *Studia Logica*, 43(1–2):131–158, 1984.

Richard Routley and Valerie Routley. Semantics of first degree entailment. *Noûs*, 6(4): 335–359, 1972.

G. Russell. Logic isn't normative. *Inquiry*, 63:371–388, 2017.

John C. Shepherdson. Negation as failure: a comparison of Clark's completed data base and Reiter's closed world assumption. *The Journal of Logic Programming*, 1(1):51–79, 1984.

John C. Shepherdson. Negation as failure, II. *The Journal of Logic Programming*, 2(3): 185–202, 1985.

Anthia Solaki, Francesco Berto, and Sonja Smets. The logic of fast and slow thinking. *Erkenntnis*, 86(3):733–762, 2019.

E. Stein. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford University Press, Oxford, 1997.

F. Steinberger. The normative status of logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Stanford, CA, 2022.

K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, Cambridge, MA, 2008.

P. C. Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3): 273–281, 1968.

H. Yama. Matching versus optimal data selection in the wason selection task. *Thinking and Resasoning*, 7:295–311, 2001.