Framing Effects

Berto, F.; Özgün, A.

DOI
10.1093/oso/9780192857491.003.0007
Publication date
2022
Document Version
Final published version
Published in
Topics of Thought

7

Framing Effects

Co-authored with Aybüke Özgün

Framing effects concern one's having different attitudes towards logically or necessarily equivalent propositions (Kahneman and Tversky 1984). Framing is, thus, connected to the hyperintensionality of thought, which we know our TSIMs to be good at modelling. However, the sort of framing effects typically investigated in cognitive science, behavioural economics, decision theory, and the social sciences at large, may benefit from a bit more specificity than the kinds of hyperintensionality the TSIMs have been put to work to model so far, in particular in the hyperintensional account of belief presented in the previous chapter.

Typical framed believers are clearly logically non-omniscient. But what *kind* of non-omniscience do they display? Section 7.1 delves into this. Specifically: such believers can have different attitudes towards intensionally equivalent φ and ψ, even if they are perfectly on top of the relevant subject matters, and, in a sense, aware of the equivalence. In order to represent this, we may need more than the plain topic-sensitivity of belief in focus in the previous chapter.

Section 7.2 introduces what we take to be the required additional ingredient. A key distinction we need in order to model typical framing, we submit, is the structural one, borrowed from cognitive psychology, between beliefs *activated* in working memory and beliefs left *inactive* in long-term memory. Few proposals in epistemic logic have featured

Topics of Thought: The Logic of Knowledge, Belief, Imagination.
Francesco Berto, Oxford University Press. © Francesco Berto 2022.
DOI: 10.1093/oso/9780192857491.003.0007

formalizations of such a distinction, whereas there is an amount of literature on the distinction between explicit and implicit belief. We discuss some of it in Section 7.3. We get to our own proposal, spelling out a formal semantics, in Section 7.4. Finally, in Section 7.5 we explore its logic. In Berto and Özgün (2021) a sound and complete axiomatization is presented; we only discuss here some notable validities and invalidities.

7.1 Framed Believers

Physicians tend to believe some lung cancer patients should get surgery with a 90% one-month survival rate. Physicians tend not to believe such patients should get surgery with a 10% first-month mortality (Kahneman 2011, 367). People will believe more in a certain economic policy when its employment rate is given than when the corresponding unemployment rate is given (Druckman 2001b). Early student registration is boosted by threatening a lateness penalty more than by promising an early bird discount (Gächter et al. 2009). A good deal of behavioural economics takes its cue from framing effects. Unlike Econs, the fully consistent agents of classical economic theory who well-order their preferences and maximize expected utility, Humans can be framed: nudged into believing different things depending on how equivalent options are presented to them (Thaler and Sunstein 2008). Framing has momentous social consequences (Plous 1993; Druckman 2001a; Levin et al. 2002; Busby et al. 2018). We need a logic of framing.

We have seen since chapter 1 that our non-omniscience is tied to different, often orthogonal, features of our cognitive apparatus. This is especially relevant here. What kind of non-omniscience is involved in framing? It cannot be tied to the *a priori/a posteriori* distinction (as in, one believes that John is John, not that John is Jack the Ripper): that the survival rate is 90% is neither more nor less *a priori* than that mortality is 10%. Nor can it be due to computational difficulties with parsing long and syntactically complex sentences ($\varphi \supset \varphi$ vs complicated tautology): either of 'The survival rate is 90%'

and 'The mortality (rate) is 10%' is just as easy to parse as the other.

It may be that the issue is not with the nature of the attitude itself either, independently of cognitive and computational limitations. We've been exploring at length the idea that knowledge or knowability ascriptions may fail full closure in chapter 4: logically astute reasoners might fail to (be positioned to) know they are no recently envatted brains although they know they have hands, etc. We saw that the jury is out on this. What we are after now, however, is belief. When the case for knowledge not being closed under entailment even for deductively unbounded reasoners is presented, their being logically astute is usually *defined* in terms of belief: they do believe all the competently deduced logical consequences of what they know (and therefore believe), based on what they know (Dretske 1970; Nozick 1981; Holliday 2015, etc.). The open issue is whether that's sufficient for the closure of their knowledge states.

Could the kind of non-omniscience displayed by agents with framed beliefs be due to a lack of concepts, as when one believes φ but not an entailed ψ because one doesn't have some notion required to grasp ψ, and perhaps specifically the topic of ψ, what it's about? We've seen that the TSIMs are especially good with this, which gets us closer to the phenomenon we're after, but perhaps not close enough. Surely human thinkers have a limited repertoire of concepts and, as a consequence, are just unable to grasp some things language and thought in the abstract can be about; but that's not what is involved in ordinary framing. Framed physicians have all the concepts needed to fully grasp both the proposition that the survival rate is 90% and the one that the mortality is 10%. In particular, they are fully on top both of the concept *survival* and of the concept *mortality* by any conceptual or semantic competence test. They may even be aware, in a sense, that the two propositions are necessarily equivalent. Still, in some other sense, they must fail to be aware, when only the former proposition gets them to believe the patients should take surgery.

What is going on, framing theorists say (Kahneman and Tversky 1984; Kahneman 2011), is that 'The mortality is

10%', but not 'The survival rate is 90%', *makes* people think about mortality. The thought that the survival rate is 90% is not about that: on the face of it, it's about survival. Survival and death are deeply connected in anyone's mind. But, cognitively limited as we are, we may not think about mortality – and much of what comes with it – when we think about survival rates, even if we have the concept *mortality* firmly in our repertoire. We leave it asleep. In order to think that the mortality is 10%, instead, we have to think about mortality, for that's what the proposition is about.

Typically framed thinkers can have different attitudes towards necessarily equivalent propositions they perfectly grasp, due to differences in what those propositions are about, even when they are perfectly on top of the relevant topics, and even when they are, in some sense, aware of the equivalence. This is not the only way agents can be framed: *qua* psychological phenomenon, framing can involve all sorts of subtle pragmatic cues and mental associations triggered by word order, emphasis, etc. But it is a typical kind of framing, we conjecture, because it has deep roots, on the one hand, in the structure of our belief system, and on the other, in the nature of its contents. An accurate logic of framing will have to represent both roots.

7.2 Working Memory, Long-Term Memory, Aboutness

To model the structural features of our belief system responsible for typical framing, we think, one should look at a key acquisition of cognitive psychology: the distinction between working and long-term memory (Eysenck and Keane 2015, part II). (To be sure – and in reply to some helpful comments of one reviewer of this book – we don't claim to have empirical evidence of deep connections between framing and that distinction. We advance this as a conjecture, which might perhaps be operationalized and tested empirically, though we 'armchair' logical modellers have no idea of how to do it.) Researchers disagree on the nature of both kinds of memory. *Qua* logical modellers, we don't want our account to be held

hostage to the next empirical discovery or consensus switch in psychological research. Luckily, we can be neutral on the more controversial issues, and just take on board the less controversial ones.

For instance, working memory (WM), which deals with the processing and short-term storage of information, is at times understood as encompassing a buffer of data at hand for the performance of cognitive tasks, plus a central executive unit: the locus of attention and cognitive control (Baddeley 1986, 2002); at times, as a plurality of modules or structures (Barsalou 1992). For our purposes, we only need to consider its most agreed-upon feature: it has limited capacity. Only a few chunks of information can be retained in WM, and only for a limited amount of time: see the views compared in Miyake and Shah (1999).

Instead, long-term memory (LTM), or the declarative part of it (Squire 1987; Schachter and Tulving 1994), is that vast knowledge base where cognitive agents store, or encode, their beliefs and knowledge about specific events (the so-called episodic memory) as well as general laws and principles (the so-called semantic memory). There's a divide in cognitive psychology, on whether WM and LTM are separate (contents are stored in LTM and retrieved from it for use in WM), or the former is just the activated part of the latter (Anderson 1983; Crowder 1993; Miyake and Shah 1999). We can be neutral on this as well.

Now, typically framed agents, we propose, can have the belief that patients should get surgery with a 90% one-month survival rate activated in their working memory, without having the intensionally equivalent belief that patients should get surgery with a 10% first-month mortality there. However, framed agents can have all the relevant information and, in particular, the concept *mortality*, in their (declarative) LTM. Let's call beliefs activated in WM *active*, and beliefs left asleep in LTM *passive*. A belief is active when it is available in WM to perform cognitive tasks with it. It is passive when it is stored, or encoded, in the agent's LTM, and left inactive there. We propose that both kinds of belief be taken as topic-sensitive. We represent them as modals in Section 7.4 below, and so they count as TSIMs.

Here are some desiderata our logic of framing should comply with. First, we evaluate ascriptions of active belief with respect to the agents' WM, and ascriptions of passive belief with respect to their LTM. Next, we may want to model realistic agents with bounded resources with respect to *both* WM *and* LTM. Psychologists contrast the limited capacity of the former with the breadth of the latter. However, neither should host all the logical consequences of what it hosts, or display an omni-inclusive conceptual repertoire. In particular, both passive and active belief must be hyperintensional: framed agents are not logically closed with respect to either.

Next, whether WM is separate from LTM, or just the activated part of the latter, no information or concept can be in WM unless it is in LTM to begin with. In particular, agents cannot have any attitude on subject matters whose concepts they simply lack. To go back to the Stalnakerian example of Section 3.3: they are as blind to them as William III was to the topic of nuclear weapons.

To get an idea of how such desiderata cooperate, consider the following two triplets of group-wise intensionally equivalent sentences:

1. $7 + 5 = 12$.
2. No three positive integers x, y and z satisfy $x^n + y^n = z^n$ for integer value of $n > 2$.
3. Extremally disconnectedness is not a hereditary property of topological spaces.
4. Triangles have three sides.
5. Bachelors are unmarried.
6. Baryons are hadrons with odd numbers of valence quarks.

(1)-(3) are necessary, of the same kind of necessity (mathematical necessity). Ditto for (4)-(6) (say, definitional necessity). Typical framed believers could find themselves in the following situation with respect to each triplet: they passively believe the first item, (1), or (4); they have the relevant information and they are on top of the basic arithmetical

or geometrical subject matter involved, so it's all stored or encoded in LTM. They are just not thinking about arithmetic, or about triangles, at the moment. They actively believe the second item, (2) or (5): they have the relevant propositional content in their WM because they are currently engaged in thoughts about diophantine equations, or John's marital status. They neither actively nor passively believe the third item, (3) or (6): they just have no idea what topological spaces are and what features they have; they have never heard about exotic notions from particle physics. This three-fold distinction isn't naturally modelled in the setting of the previous chapter (compare the examples in Section 6.1).

Before we get to our own proposal to model agents of this kind, in the next section we briefly discuss some hyperintensional epistemic logics for non-logically omniscient agents already on the market, to see to what extent they *could* be used to represent framing.

7.3 Explicit and Implicit

As far as we know, few epistemic logics have aimed at directly representing the difference between WM and LTM. One distinction which may look *prima facie* similar is the one between *explicit* and *implicit* knowledge and belief, found in awareness logics developed with an eye on the logical omniscience problem (Fagin and Halpern 1988; Van Benthem and Velázquez-Quesada 2010; Velázquez-Quesada 2014): we briefly introduced and discussed them in Section 3.3. Because being unaware of φ is usually understood as not having φ present in the mind, or not thinking about φ (Schipper 2015, 79-80), the awareness approach seems especially suitable to model framing.

Remember how awareness is typically represented syntactically: one is aware of φ when φ belongs to a set of formulas, $\mathcal{A}$, the agent's awareness set. Implicit knowledge or belief are dealt with via normal Hintikka-style modal operators, whereas the corresponding explicit attitudes are defined as the combination of the implicit ones with awareness: one

explicitly knows or believes that φ when one knows or believes it implicitly and φ is in the awareness set.

We mentioned in Section 3.3 that the view has been claimed to mix syntax and semantics, essentially imposing a syntactic filter over a standard Hintikkan semantics (Konolige 1986). Resorting to syntax, however, allows very fine-grained distinctions: if any bunch of sentences can serve as the awareness set $\mathcal{A}$, explicit attitudes obey no non-trivial logical closure properties. Syntactic approaches representing bodies of knowledge/belief/awareness as plain sets of sentences have then been criticized for being *too* fine-grained (Levesque 1984, 199-201). For the purposes of modelling the typical framing effects we're after, they are an overkill.

Here's why: a framed agent who actively believes $\varphi \wedge \psi$ should actively believe $\psi \wedge \varphi$, and should actively believe φ, if our topic-sensitive view of propositional content is right. That John is tall and handsome and that John is handsome and tall are intensionally equivalent propositions, and the agent who actively believes either is already thinking about the other's topic – because it is the same topic, say, John's height and looks. That John is tall and handsome entails that John is tall, and one who actively believes the former is already thinking about the topic of the latter, as it is part of that of the former. Such mereological relations between the contents of thoughts, which have been at centre stage for much of this book, are lost in a plain awareness setting. (This doesn't rule out, we think, that plain syntactic awareness approaches may be useful in modelling some specific kinds of framing, e.g., the presentation order effects discussed in Section 3.2. As conjectured there, these may to be tied to issues with parsing the syntax of sentences.)

Nor do implicit attitudes neatly map to passive belief as implemented in LTM. Because logics featuring the explicit/implicit distinction usually take the implicit attitude as a normal Hintikkan modality, the attitude displays full logical omniscience: the agent implicitly knows or believes all logical truths, and all logical consequences of what it knows or believes. The agent has no awareness or conceptual limitations there: it is simply on top of all the relevant propositions. But, as we have remarked, LTM is not like that.

If we want to model agents who don't possess all concepts, and don't have all the logical consequences of their passive beliefs stored or encoded in LTM, passive belief should be hyperintensional, too.

Balbiani et al. (2019) present one of the few logical works with the stated aim of modelling the WM/LTM distinction. It's a powerful framework in the tradition of Dynamic Epistemic Logic, modelling the processes through which a non-omniscient agent forms its beliefs via operations of perception and inference in WM, and can store and retrieve them from LTM. Their language has an operator for explicit belief, tied to WM, and one expressing background knowledge, tied to LTM. The latter is a normal modality, and so faces the same issue as implicit knowledge in the awareness setting: the agent is logically omniscient with respect to its background knowledge.

What's more worrying for the prospects of applying the logic to framingis that explicit belief gets a Scott-Montague neighbourhood semantics (Scott 1970; Pacuit 2017): one explicitly believes that φ when φ's truth set is in the relevant neighbourhood set. We talked of the neighbourhood approach in Section 6.1, where we mentioned that it gives weak non-normal modal logics capable of breaking a number of logical closure features for their operators. In particular, one can explicitly believe a conjunction without explicitly believing the conjuncts, which, we argued above, is not good. This overkill can be fixed by adding conditions – specifically, one could close the neighbourhoods under supersets for $\wedge$-elimination: see Pacuit (2017), 81.

However, there's still the more problematic underkill we flagged in Section 6.1: even in the basic neighbourhood setting, when φ and ψ are assigned the same set of worlds as their (thin) proposition, they will be in the same sets of neighbourhoods. Thus, explicit belief in either will automatically entail explicit belief in the other. This is exactly what shouldn't happen if we want to capture framing for explicit beliefs. As we also mentioned in that section, one can play with the addition of (mathematically, logically, etc.) impossible worlds to make neighbourhood semantics more fine-grained. But the topic-sensitive approach may be better

positioned to capture how the subjects of typical framing fail to (actively) think about one of two topics driving a wedge between intensional equivalents. Thus, we now move on to our own proposal and start making things formally precise.

7.4 Topic-Sensitive Active and Passive Belief

Our language $\mathcal{L}$ for this chapter will have, besides the countable set $\mathcal{L}_{AT}$ of atomic formulas p, q, r $(p_1, p_2...)$, negation $\neg$, conjunction $\wedge$, disjunction $\vee$, the box of necessity $\Box$, and two belief operators, B_A and B_P. The well-formed formulas are the items in $\mathcal{L}_{AT}$ and, if φ and ψ are formulas, so are the following:

$$\neg\varphi \mid (\varphi \wedge \psi) \mid (\varphi \vee \psi) \mid \Box\varphi \mid B_A\varphi \mid B_P\varphi$$

As usual, we often omit outermost brackets and we identify $\mathcal{L}$ with the set of its well-formed formulas. Read the box as a normal epistemic or *a priori* modality (flag this: we may then see the worlds of the coming semantics as epistemically possible ones, rather than absolutely or broadly metaphysically possible ones; given that the modal is a normal one, the former will not differ that much from the latter anyway – they may, e.g., falsify narrowly metaphysically necessary claims like 'Hesperus is Phosphorus' or 'Water is H2O'); read '$B_A\varphi$' as 'One actively believes that φ', '$B_P\varphi$' as 'One passively believes that φ'. When we say something that applies to both active and passive belief, we use 'B_*'. It will come in handy to have a $\top := p \vee \neg p$ (this abbreviates a specific tautology; it is not to be confused, thus, with the $\top$ of Section 6.4) and a $\bot := \neg\top$. Again, '$\mathfrak{At}\varphi$' stands for the set of atomic formulas occurring in φ.

A *frame* for $\mathcal{L}$ is a tuple $\mathfrak{F} = \langle W, \mathcal{O}, \mathcal{T}, \oplus, t\rangle$, where W is our non-empty set of possible worlds, $\mathcal{T}$ is our non-empty set of topics, $\oplus$ is topic fusion (with topic parthood, $\leq$, defined from it as usual). The new bit is $\mathcal{O}$, a non-empty, finite subset of $P(W)$ such that $\mathcal{O} \neq \{\emptyset\}$: each non-empty $O \in \mathcal{O}$ represents the informational content of a *memory cell* (we'll come to what this is in a second).

The topic function now is $t : \mathcal{L}_{AT} \cup \mathcal{O} \to \mathcal{T} \cup P(\mathcal{T})$. It assigns a topic to each atomic formula and a non-empty, finite set of topics to each item in $\mathcal{O}$: $t(p) \in \mathcal{T}$ for all $p \in \mathcal{L}_{AT}$, and $t(O) \in P(\mathcal{T})$ is non-empty and finite for all $O \in \mathcal{O}$. Then, topics are assigned to the whole of $\mathcal{L}$ the usual way, namely with $t(\varphi) = \oplus\mathfrak{At}\varphi$, to ensure topic-transparency.

Here's what the model represents: the agent's belief system is composed of memory cells. These are chunks of LTM which can be put into (or, if one prefers, activated as) WM, that is, made available for actions of cognitive processing. A memory cell is represented by an indexed set, O_x, where $\emptyset \neq O \in \mathcal{O}$ and $x \in t(O)$. O_x is made of informational content O and topic x. Memory cells are, thus, topic-sensitive: when one is in (or activated as) WM, the agent is actively thinking about its subject matter, and has its informational content available for processing. $t(O)$ and $\mathcal{O}$ are assumed to be finite, to represent cognitive agents that can only have finitely many memory cells.

Every $O \in \mathcal{O}$ is assigned a set of topics, rather than a single topic, in order to capture the idea that the same informational content can be associated with different topics. Take our triplet of intensionally equivalent, topic-diverging sentences (1), (2), and (3) in Section 7.2. Intensional equivalence means that they have the same bunch of worlds as their truth set. Call it S. Let the topics be x, y, and z, respectively. Each of S_x, S_y, and S_z can make for a distinct memory cell, differing from the others in topic but not in informational content.

The agent's LTM is defined as:

$$LTM := (\bigcap \mathcal{O})_{\oplus(\bigcup_{O\in\mathcal{O}} t(O))}$$

The information stored or encoded in LTM is the information available in all memory cells, taken together. The topic of LTM is the fusion of those of all memory cells: the total repertoire of subject matters the agent has grasped. To simplify the notation, we set $\bigcap \mathcal{O} := O^{\cap}$ and $\oplus(\bigcup_{O\in\mathcal{O}} t(O)) := \mathfrak{b}$. Then the LTM of the agent is $O^{\cap}_{\mathfrak{b}}$, which features the 'total topic' the agent is on top of. Notice that $\mathfrak{b}$ is guaranteed to be in $\mathcal{T}$, since $\bigcup_{O\in\mathcal{O}} t(O)$ is finite.

LTM is larger than any single memory cell which can be activated as, or put into, WM, with respect to both

information and topic. The agent passively believes, i.e., has in LTM, way more than it can actively believe, i.e., activate and process in WM: the latter has quite limited capacity compared to LTM, as cognitive psychology has taught us.

Next, a *model* $\mathfrak{M} = \langle W, \mathcal{O}, \mathcal{T}, \oplus, t, \Vdash \rangle$ is a frame with an interpretation $\Vdash$, which works differently from what we've seen in previous chapters: we now evaluate formulas with respect to world-memory pairs, $\langle w, O_x \rangle$, with $w \in W$ representing the actual world, and O_x a memory cell. The working memory WM is just the designated world-memory cell with respect to which we evaluate formulas. We denote the set of all world-memory pairs of model $\mathfrak{M}$ as $\mathcal{P}(\mathfrak{M})$ ('$\mathcal{P}$' is for 'pair', not the power set operation). The interpretation relates such pairs to atomic formulas: we read '$\langle w, O_x \rangle \Vdash p$' as saying that p holds at $\langle w, O_x \rangle$, '$\langle w, O_x \rangle \nVdash p$' as: $\sim \langle w, O_x \rangle \Vdash p$. This is extended to all formulas of $\mathcal{L}$ thus:

(S$\neg$) $\langle w, O_x \rangle \Vdash \neg\varphi \Leftrightarrow \langle w, O_x \rangle \nVdash \varphi$

(S$\wedge$) $\langle w, O_x \rangle \Vdash \varphi \wedge \psi \Leftrightarrow \langle w, O_x \rangle \Vdash \varphi$ & $\langle w, O_x \rangle \Vdash \psi$

(S$\vee$) $\langle w, O_x \rangle \Vdash \varphi \vee \psi \Leftrightarrow \langle w, O_x \rangle \Vdash \varphi$ *or* $\langle w, O_x \rangle \Vdash \psi$

(S$\Box$) $\langle w, O_x \rangle \Vdash \Box\varphi \Leftrightarrow W \subseteq |\varphi|^{O_x}$

(SB_A) $\langle w, O_x \rangle \Vdash B_A\varphi \Leftrightarrow$ [1] $O \subseteq |\varphi|^{O_x}$ & [2] $t(\varphi) \leq x$

(SB_P) $\langle w, O_x \rangle \Vdash B_P\varphi \Leftrightarrow$ [1] $O^{\cap} \subseteq |\varphi|^{O_x}$ & [2] $t(\varphi) \leq \mathfrak{b}$

where $|\varphi|^{O_x} = \{w \in W | \langle w, O_x \rangle \Vdash \varphi\}$.

Both active and passive belief are topic-sensitive and get TSIM-style, two-component truth conditions. For $B_*\varphi$ to come out true at $\langle w, O_x \rangle$, we ask for two things to happen: [1] φ must be entailed by the information O in WM for active belief, and by the information $O^{\cap}$ in LTM for passive belief; and [2] the topic of φ must be included in the topic x activated in WM, for active belief, and in the overall LTM topic $\mathfrak{b}$ the agent is on top of, for passive belief.

Only the truth value of an ascription of *active* belief depends on the chosen O_x.[1] However, the agent can believe

[1] Given a model $\mathfrak{M} = \langle W, \mathcal{O}, \mathcal{T}, \oplus, t, \Vdash \rangle$, $w \in W$, two world-memory pairs $(w, O_x), (w, U_y) \in \mathcal{P}(\mathfrak{M})$, and $\varphi \in \mathcal{L}$ such that φ does not have any occurrences of B_A, we have: $\langle w, O_x \rangle \Vdash \varphi \Leftrightarrow \langle w, U_y \rangle \Vdash \varphi$.

φ with respect to one memory cell without believing the same content with respect to another one. That is, given a model $\mathfrak{M} = \langle W, \mathcal{O}, \mathcal{T}, \oplus, t, \Vdash \rangle$ and two world-memory pairs $\langle w, O_x \rangle, \langle w, U_y \rangle \in \mathcal{P}(\mathfrak{M})$, it could be that $\langle w, O_x \rangle \Vdash B_A\varphi$ and $\langle w, U_y \rangle \nVdash B_A\varphi$ for some $\varphi \in \mathcal{L}$, as shown in the Sample Model in the footnote.[2]

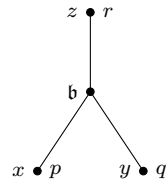
Finally, valid entailment is truth preservation at all world-memory pairs of all models. With Σ a set of formulas:

$\Sigma \vDash \psi \Leftrightarrow$ in all models $\mathfrak{M} = \langle W, \mathcal{O}, \mathcal{T}, \oplus, t, \Vdash \rangle$ and for all $\langle w, O_x \rangle \in \mathcal{P}(\mathfrak{M})$: $\langle w, O_x \rangle \Vdash \varphi$ for all $\varphi \in \Sigma \Rightarrow \langle w, O_x \rangle \Vdash \psi$

For single-premise entailment, we write $\varphi \vDash \psi$ for $\{\varphi\} \vDash \psi$. Validity for formulas, $\vDash \varphi$, truth at all world-memory pairs of all models, is $\emptyset \vDash \varphi$, entailment by the empty set of premises.

We'll make use of the abbreviation $\bar{\varphi} := \bigwedge_{p \in \mathfrak{At}\varphi}(p \vee \neg p)$. This will play a role in formalizing validities and invalidities.[3]

[2]Let $\mathfrak{M} = \langle \{w, w_1, w_2\}, \{O, U\}, \{x, \mathfrak{b}, y, z\}, \oplus, t, \Vdash \rangle$ such that $O = \{w, w_1\}$, $U = \{w, w_2\}$, and $\langle \{x, \mathfrak{b}, y, z\}, \oplus \rangle$ constitutes the join-semilattice in this figure:



The dots are topics and the lines represent topic-inclusion relations, going upwards. So for t we have $t(p) = x, t(q) = y$, $t(r) = z$. As for $\Vdash$, p and q's truth set is $\{w, w_1\}$, r's truth set is $\{w, w_2\}$. Then $(w, O_x) \Vdash B_Ap$ since $O \subseteq \{w, w_1\}$ and $t(p) \leq x$. However, $(w, O_y) \nVdash B_Ap$ since $t(p) \not\leq y$, that is, the agent does not have the subject matter of p in working memory O_y. Similarly, we also have, e.g., $(w, U_y) \nVdash B_Ap$ for two reasons: (1) $U \not\subseteq \{w, w_1\}$ and (2) $t(p) \not\leq y$, that is, the informational content of U_y does not eliminate all non-p possibilities and the subject matter of p is not part of the subject matter of working memory U_y, respectively.

[3]In order to have a unique definition of each $\bar{\varphi}$, we set the convention that elements of $\mathfrak{At}\varphi$ occur in $\bigwedge_{p \in \mathfrak{At}\varphi}(p \vee \neg p)$ from left-to-right in the order they are enumerated in $\mathcal{L}_{AT} = \{p_1, p_2, \dots\}$. For example, for $\varphi := B_*(p_{10} \to p_2) \vee \Box p_7$, $\bar{\varphi}$ is $(p_2 \vee \neg p_2) \wedge (p_7 \vee \neg p_7) \wedge (p_{10} \vee \neg p_{10})$, and not $(p_{10} \vee \neg p_{10}) \wedge (p_7 \vee \neg p_7) \wedge (p_2 \vee \neg p_2)$ or $(p_7 \vee \neg p_7) \wedge (p_{10} \vee \neg p_{10}) \wedge (p_2 \vee \neg p_2)$ etc. This convention will eventually not matter since our logic cannot differentiate two conjunctions of different order: $\varphi \wedge \psi$ is provably and semantically equivalent to $\psi \wedge \varphi$.

Given a model $\mathfrak{M} = \langle W, \mathcal{O}, \mathcal{T}, \oplus, t, \Vdash \rangle$, it is easy to see that $\bar{\varphi}$ is true at every world-memory pair in $\mathcal{P}(\mathfrak{M})$ and $\mathfrak{At}\bar{\varphi} = \mathfrak{At}\varphi$ for any $\varphi \in \mathcal{L}$. This trick allows us to talk inside the language about what topics the agent is actively thinking about in WM, and what topics the agent has grasped and stored in LTM. Formulas of the form $B_A\bar{\varphi}$ ($\neg B_A\bar{\varphi}$) express within $\mathcal{L}$ statements such as 'The agent has (does not have) the subject matter of φ in WM'.[4] Similarly, formulas of the form $B_P\bar{\varphi}$ ($\neg B_P\bar{\varphi}$) express within $\mathcal{L}$ statements such as 'The agent has (does not have) the subject matter of φ in LTM'.

Our semantics is a variation on the *subset space semantics* of Moss and Parikh (1992), in that the component $\langle W, \mathcal{O} \rangle$ of our frames is a subset space (a pair of a set and a selection of its subsets) and we evaluate sentences not at worlds but at world-set pairs. Subset space semantics was originally designed to model an evidence-based notion of absolutely certain knowledge and epistemic effort. The evaluation pairs of the form $\langle w, O \rangle$ within this framework obey the constraint that $w \in O$ (for knowledge is veridical) and are often called 'epistemic scenarios'. O represents the agent's current truthful evidence.

Our framework comes with a distinct formalism, however, and a different interpretation of a subset space model's components. We focus on belief rather than knowledge, so the evaluation pairs are tailored accordingly: as belief is not factive, a memory cell $\langle w, O_x \rangle$ does not have to meet the constraint $w \in O$. More importantly, our subset spaces and the corresponding evaluation pairs are endowed with topics. This makes the resulting logic of belief hyperintensional, as opposed to the intensional epistemic logics of the traditional subset space semantics (Moss and Parikh 1992; Dabrowski et al. 1996; Weiss and Parikh 2002).

7.5 The Logic of Framing

In Berto and Özgün (2021), we come up with a sound and complete axiomatization L for the logic of framed belief over

[4]Notice that $\langle w, O_x \rangle \Vdash B_A\bar{\varphi} \Leftrightarrow O \subseteq |\bar{\varphi}|^{O_x}$ & $t(\bar{\varphi}) \leq x \Leftrightarrow O \subseteq W$ & $t(\varphi) \leq x \Leftrightarrow t(\varphi) \leq x$.

$\mathcal{L}$. The axioms are:

(CPL) All classical tautologies and Modus Ponens;

($S5_\Box$) S5 axioms and rules for $\Box$;

(I) Axioms for B_*, with $* \in \{A, P\}$:

(C_{B_*}) $B_*(\varphi \wedge \psi) \equiv (B_*\varphi \wedge B_*\psi)$
($\mathsf{Ax1}_{B_*}$) $B_*\varphi \supset B_*\overline{\varphi}$
($\mathsf{Ax2}_{B_*}$) $(\Box(\varphi \supset \psi) \wedge B_*\varphi \wedge B_*\overline{\psi}) \supset B_*\psi$
($\mathsf{Ax3}_{B_*}$) $B_*\varphi \supset \Box B_*\varphi$

(II) Axioms for B_A:

(D_{B_A}) $B_A\varphi \supset \neg B_A\neg\varphi$

(III) Axioms connecting B_A and B_P:

(Inc) $B_A\varphi \supset B_P\varphi$

The notion of derivation, denoted by $\vdash$, in $\mathfrak{L}$ is defined as usual. Thus, $\vdash \varphi$ means φ is a theorem of $\mathfrak{L}$. $\mathfrak{L}$ is a sound and complete axiomatization of $\mathcal{L}$ with respect to the class of models given above: for every $\varphi \in \mathcal{L}$, $\vdash \varphi$ if and only if $\vDash \varphi$ (see the appendix to our paper for the proof).

The axioms in Group I give general closure features of belief, both active and passive, for our framed agents. C_{B_*} ensures that beliefs are fully Conjunctive, as usual for our TSIMs and as per the defence in Section 3.2: one who believes, either actively or passively, that John is tall and handsome, believes both that John is tall and that John is handsome, and vice versa. $\mathsf{Ax1}_{B_*}$ captures, as desired, the topic-sensitivity of belief: one can actively believe φ only if one is actively thinking about the relevant topic in WM; one can passively believe φ only if one has concepts for the relevant topic stored in LTM. $\mathsf{Ax2}_{B_*}$ states a limited deductive closure principle for both active and passive belief: if ψ follows from φ *a priori*, and one believes φ, *and* one is on top of the subject matter of ψ, then one does believe ψ. $\mathsf{Ax3}_{B_*}$ has it that beliefs are not world-relative.

In Group II, D_{B_A} states a consistency principle for active belief: one who has φ in WM will not also have $\neg\varphi$ there. Notice that this does not hold for passive belief: our framed agent may have all sorts of inconsistent beliefs stored or encoded in its LTM. They can stay there insofar as one does not think about them all together. This makes for a very realistic modelling: isn't this the way we are, for the most part? We are quite inconsistent in the beliefs we hold – provided the inconsistencies remain stored in our long-term memory, shielded from the focus of our attention.

As for Group III, the Inc principle bridges active and passive belief. It guarantees, as desired, that whatever is activated in WM be available in LTM to begin with.

As always with our TSIMs, just as important as validities are the invalidities involving them, as they display the precise sort of non-omniscience our framed agents instantiate. We discuss a few prominent invalidities:

1. From φ, infer $B_*\varphi$ [Omniscience Rule]
2. $\Box\varphi \supset B_*\varphi$ [*A Priori* Omniscience]
3. $(\Box(\varphi \supset \psi) \wedge B_*\varphi) \supset B_*\psi$ [Closure under *A Priori* Implication]
4. $\neg B_*\varphi \supset B_*\neg B_*\varphi$ [Negative Introspection]
5. From $\varphi \equiv \psi$, infer $B_*\varphi \equiv B_*\psi$ [Framing-A]
6. $(B_A\varphi \wedge B_P(\varphi \equiv \psi)) \supset B_A\psi$ [Framing-B]
7. From $\varphi \equiv \psi$, infer $(B_A\varphi \wedge B_P\overline{\psi}) \supset B_A\psi$ [Framing-C][5]

[5] *Countermodel*: take our Sample Model from footnote 2 above. We have (1) and (2) invalid since $\vDash r \vee \neg r$ (therefore also $(w, O_x) \Vdash \Box(r \vee \neg r)$), but $(w, O_x) \nVdash B_A(r \vee \neg r)$ (since $t(r) \not\leq x$) and $(w, O_x) \nVdash B_P(r \vee \neg r)$ (since $t(r) \not\leq \mathfrak{b}$). For (3), take $\varphi := p$ and $\psi := r \vee \neg r$: $(w, O_x) \Vdash \Box(p \to (r \vee \neg r))$, $(w, O_x) \Vdash B_A p$, and $(w, O_x) \Vdash B_P p$, however, $(w, O_x) \nVdash B_A(r \vee \neg r)$, and $(w, O_x) \nVdash B_P(r \vee \neg r)$ as shown above. For (4), take $\varphi := r$: world-memory pair (w, O_x) falsifies it for B_A since $t(r) = t(\neg B_A r) \not\leq x$ and falsifies it for B_P since $t(r) = t(\neg B_P r) \not\leq \mathfrak{b}$. For (5), take $\varphi := p \vee \neg p$ and $\psi := r \vee \neg r$, and (w, O_x) falsifies the principle. For (6), take $\varphi := p$ and $\psi := q$: $(w, O_x) \Vdash \mathsf{B}_A p$ and $(w, O_x) \Vdash B_P(p \leftrightarrow q)$, but $(w, O_x) \nVdash B_A q$ (since $t(q) \not\leq x$). For (7), take $\varphi := p \vee \neg p$ and $\psi := q \vee \neg q$, and observe that (w, O_x) falsifies the principle.

The failure of (1)-(3) tells us that our agents don't believe all (*a priori*) truths and that their beliefs are not closed under *a priori* implication. (4) says that they lack the wisdom of negative introspection: they can fail to believe that they don't believe something.

The last three invalidities, (5)-(7), crucially capture the typical framing we were after: Framing-A guarantees that agents can have different attitudes towards intensionally equivalent formulas. Framing-B says that one can have the belief that φ (e.g., patients should get surgery with a 90% one-month survival rate) activated in WM, without having the belief that ψ (patients should get surgery with a 10% first-month mortality) there, even when one *does* have their equivalence in one's LTM. In this sense, one is aware: one is on top of all the relevant concepts and does believe that either is true iff the other is. But all of this is left asleep in LTM. In this other sense, one is not aware: one is just not thinking about it. Framing-C says that one's actively believing φ does not imply that one actively believes ψ, even when the two are equivalent and one has the subject matter of ψ in one's LTM.

Here's something the logic does not capture (an admission prompted by a remark by one reviewer of this book): the positive/negative *polarity* displayed by pairs of claims involved in typical cases of framing – death and survival rates, penalties and discounts, etc. The reviewer graciously granted that perhaps a logic of framing is not supposed to do this in a general setting. We hope so, for right now, we don't know how to tweak ours so that it does.

We close the chapter by mentioning two directions of further investigation: first, both active and passive belief TSIMs are plain, categorical forms of belief. It may be interesting to expand the language and formal semantics so that they include conditional, topic-sensitive active and passive belief, as per the two-place TSIMs we explored in previous chapters.

Second, working memory is properly so-called in cognitive psychology because it is the locus of cognitive activity: beliefs are in there in order to be manipulated, expanded, revised via operations of combination, deduction, etc. Another direction of expansion may then feature the addition to our language

of topic-sensitive dynamic operators in the style of Dynamic Epistemic Logic, perhaps as per the route summarized in Section 6.4. This would allow one to properly model how agents operate on their active beliefs in the light of new incoming information, before storing the results in LTM.

7.6 Chapter Summary

This chapter has introduced two kinds of one-place TSIMs representing, respectively, belief activated in working memory, and belief left passively stored in long-term memory. The distinction between the two sorts of belief has been shown to model a typical form of the well-known framing effect, whereby people can have different attitudes towards logically or necessarily equivalent propositions. The chapter has introduced a semantics for active and passive topic-sensitive belief to represent, and reason about, agents whose belief states can be subject to framing effects. The analysis of framing has called for a precise characterization of the sense in which framed agents are logically non-omniscient, given that they can believe exactly one of two intensionally equivalent propositions even when they are fully on top of the relevant subject matters and, in a 'dormant' sense, they are aware of the equivalence.