This is a pre-publication draft version of:

Dołęga, Krzysztof (in press). Models of introspection vs introspective devices: Testing the research programme for possible forms of introspection. For a special issue of *Journal of Consciousness Studies* edited by François Kammerer and Keith Frankish on their target article “What forms could introspective systems take? A research programme.”

Models of introspection vs introspective devices: Testing the research programme for possible forms of introspection

Krzysztof Dołęga
Université libre de Bruxelles

Abstract: The introspective devices framework proposed by Kammerer and Frankish (2023) offers an attractive conceptual tool for evaluating and developing accounts of introspection. However, the framework assumes that different views about the nature of introspection can be easily evaluated against a set of common criteria. In this paper, I set out to test this assumption by analyzing two formal models of introspection using the introspective device framework. The question I aim to answer is not only whether models developed outside of philosophy can be successfully evaluated against the set of conceptual criteria proposed by Kammerer and Frankish, but also whether this kind of evaluation can reveal some limitations inherent to the framework.

Kammerer and Frankish’s (2023, henceforth K&F) conceptual framework for mapping out possible forms of introspection is a timely and much needed theoretical development that might just reinvigorate debates on the nature of this ubiquitous yet elusive mental phenomenon. Unlike much of recent literature, K&F’s framework does not focus on the question of introspective reliability. Instead, the authors propose to evaluate different theories along the dimensions of introspection’s assumed (in)directness, (in)flexibility, and the format of the mental states that play the role of its output. By locating different theories along these three dimensions, the framework allows us to not only compare, but also to clarify existing positions about introspection. KF’s proposal also holds promise of fostering the development of new theories of non-human introspection within comparative studies on cognition and AI.

However, despite its novelty and promise of theoretical advancements, K&F's proposal is just that — a proposal of a research program. This, of course, is not a deficiency, after all every research program needs to start somewhere. Nevertheless, the programmatic nature of K&F's conceptual framework invites an important question about the ways in which it could be applied in practice. K&F seem to focus solely on the task of delineating the range of possible introspective devices or faculties, but their approach says little about how particular proposals should be evaluated against their chosen criteria. This issue is especially pertinent when considering the empirical future of the proposed framework, since not all hypotheses about introspection are spelled out in a way that will allow them to be easily located within K&F's conceptual structure.

The aim of the present article is to address the above worries by testing K&F's framework against two formal models which have been proposed to operationalize introspection — signal detection theory (or SDT) and metacognitive networks (MNs). The point of such evaluation is not just to assess where said models fall along K&F's chosen dimensions of introspection, but also to see the extent to which the framework can successfully drive model development and refinement. In other words, the question this paper intends to answer is not just whether the framework can be applied to non-philosophical models of introspection, but also whether such applications can yield insights informative for development and refinement of future models of introspective faculties.

1. Signal Detection Theory as a model of introspection

SDT is a formal model that has been initially developed for the purpose of separating signal from noise in radar applications but was later adopted to model perceptual processes in vision science (Tanner & Swets, 1954) and psychophysics (Green & Swets, 1966), where it is used to this day.

The main assumption of SDT is that, in order to perceive a stimulus, the observer needs to successfully distinguish between the internal perceptual response to the presence or absence of a stimulus (or more simply the *internal signal*) despite the ever-present uncertainty caused by factors related to internal reliability and external conditions (or more simply *noise*). Thus, SDT casts perception as a result of the observer's *perceptual sensitivity* to their internal signal and a *response bias* in how they classify their internal perceptual responses in order to separate them from noise.

To illustrate the formal details of SDT, consider a simple example of an experiment in which the subject is presented with a brief flash of light on half of the trials and must indicate whether or not they have seen the flash. In this case, the space of possible discriminations will involve two stimulus classes - one for the absences of light and one for its presence - call them *C1* and *C2* respectively. Given that there will always be variability in internal responses, the two stimulus classes are usually assumed to take the form of normal probability distributions of equal variance. The subject's perceptual (or detection) sensitivity is measured as the distance between the means of the two distributions denoted as *d'*. Low perceptual sensitivity thus corresponds with a high degree of overlap between the distributions, indicating that the subject is poor at distinguishing between proper internal responses and noise. Considering these assumptions, the task of the subject is to make a perceptual decision as to which of the two distribution their internal perceptual response *p* has been drawn from. To do this, they need to set a criterion *c* that reflects their subjective strategy for categorizing different internal responses (see Figure 1).

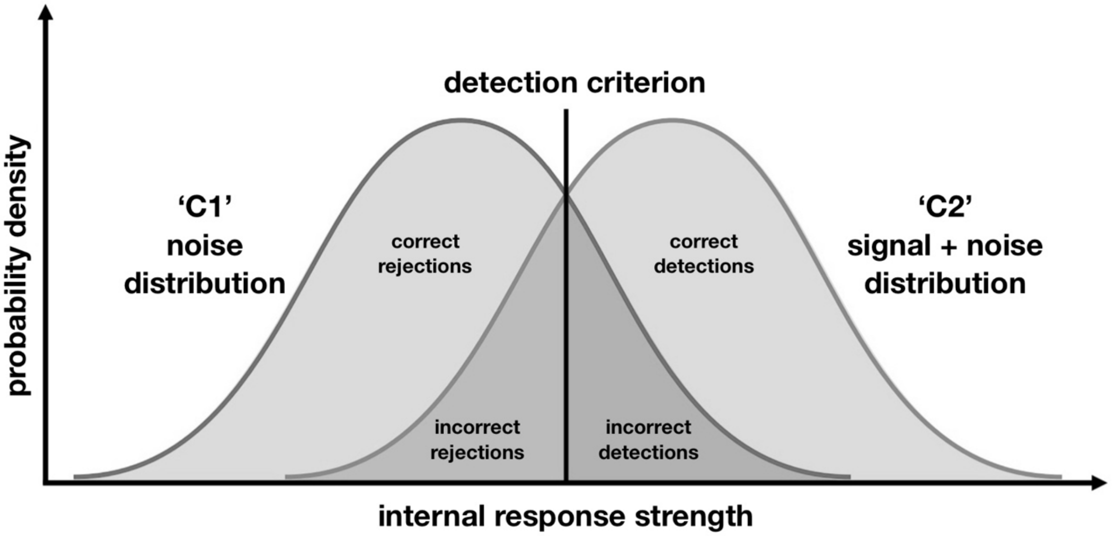

Figure 1: A typical SDT model in which perception is operationalized as a decision about internal evidence. Subjects need to classify a sample internal response as resulting from noise (distribution *C1* on the left) or as evidence of an internal signal indicating the presence of a stimulus (distribution *C2* on the right) by adopting a decision criterion (represented as solid vertical line). Any sample that falls below the criterion will be classified as a result of noise, whereas any sample falling above it will be classified as a signal indicative of the stimulus. Notice, that the chance of incorrect classifications depends not only on the subject's choice of detection criterion, but also on their sensitivity (i.e., the degree of overlap between the distributions), the amount of noise, as well as the strength of the internal response (i.e., its location along the x-axis). Adapted from Morales (forthcoming).

SDT, as outlined here, has been widely adopted in research on consciousness as a way of operationalizing perceptual awareness in humans (Lau, 2008) and other species (see e.g., Nieder, Wagener & Rinnert, 2020), though there is an ongoing debate about which of the many measures applied in this kind of analysis should be taken as reflective of the presence of conscious perception (Timmermans & Cleeremans, 2015). More importantly for present purposes, SDT has recently been proposed as a theory of introspection (Morales, *forthcoming*).

As the name suggests, the Introspective Signal Detection Theory (or iSDT) proposes that introspection is just another form of signal detection. In the same way that the strength of internal perceptual signals can vary in response to external stimuli, conscious experiences can vary with regard to their vividness or intensity, which, in turn, will have an impact on how reliably they can be introspected. Morales (forthcoming, p. 20) labels the degree of an experience's phenomenal intensity or the prominence it has within one's conscious field as that experience's *mental strength*. However, as in SDT, the presence of noise in the cognitive system means that the strength of responses will not always correlate with what is really happening; “a weak experience could occasionally generate a strong introspective response, or a strong experience could occasionally generate a weak introspective response” (ibid, p. 23).

Thus, the core proposal of iSDT is that, analogously to the application of signal detection to perception, introspective judgements depend on the subject’s introspective sensitivity to differences between experiences (measured again by *d’*) and their introspective bias (again denoted by *c*) determining how well they can distinguish between experiences of varying mental strengths. As Morales himself explains, iSDT presents a picture on which:

> *Introspecting is modeled as an introspector deciding whether an internal introspective response i was generated by a conscious-experience class C1 (for example, ‘pain absent’, ‘burning pain’, etc.) or C2 (for example, ‘pain present’, ‘stabbing pain’, etc.). The introspective response corresponds to the strength of the introspective evidence, in turn modulated by the intensity of the conscious experience (its mental strength). Repeated experiences of the same class produce introspective responses with different values due to ever-present noise of different sorts. The values of the introspective response are distributed across a decision axis. The introspective response i in any given case can be thought of as being drawn from either a noise or a signal-plus-noise distribution […] (Morales, forthcoming, p. 23).*

An important consequence of Morales’ presentation is that iSDT is a theory of *phenomenal introspection*, since it seems to assume that introspection will always be intimately linked to consciousness. However, the possibility of misclassifying non-conscious states or internal noise as experiences means that, even though introspection is conceptualized as a conscious faculty, its objects do not have to be conscious prior to being targeted.[1]

[1] I have omitted some details of iSDT for the sake of brevity. Perhaps the most interesting aspect that has been left out is subjective confidence, which takes the form of confidence criteria that separate regions of low confidence surrounding the main decision criterion from regions of higher confidence that are located further towards the peripheries of the decision space. Importantly, Morales does *not* explain *how* such confidence criteria should be chosen.

2. Metacognitive Networks as a model of introspection

One of the controversies surrounding the use of SDT in psychology is whether subjects' responses, given their decision biases (i.e., the values of *c* they choose), are a reliable measure of stimulus awareness, which might not be an all-or-nothing phenomenon. Subjective metacognitive test of awareness, such as Post-Decision Wagering (or PDW, see e.g. Persaud, McLeod & Cowey, 2007) were introduced to alleviate this worry by testing not only whether subjects successfully register information about some state of affairs (e.g., by successfully discriminating stimuli in a signal detection task), but also how confident they are in their judgements (i.e, testing the subjective uncertainty about their awareness of a stimulus). PDW aims to indirectly measure subjects' confidence in their decision process by asking them to bet on the correctness of their responses in an experimental task after they have made the decision. Assuming that subjects will try to maximize their earnings, high bets will correspond to a high degree of confidence and low bets will correspond to decisions burdened with a high degree of uncertainty and, therefore, low confidence.

MNs are a type of artificial multi-level neural networks aimed at formally modeling the acquisition and function of meta-knowledge in tests that utilize PDW to measure subjects' metacognitive awareness of their decisions (Pasquali, Timmermans & Cleeremans, 2010). An MN, of the type considered here, is made up of two interconnected networks. The first-order system is a three-layer feedforward auto-associative network (meaning its input and output layers are of the same size) that uses a winner-take-all algorithm on the output layer (meaning one unit with the highest activation is selected to determine the response) and is trained using backpropagation. The network effectively learns to solve some first-order task, e.g., perceptual signal detection, by discriminating between presence and absence of a visual stimulus. The second-order network is made up of a layer of hidden units and an output layer consisting of two nodes that are used to place low or high wagers about the first-order network's performance. This network also utilizes a winner-takes-all algorithm on its output, but its hidden units act as a matrix of comparators that compute the difference between corresponding pairs of input and output units in the first-order network. As Pasquali and colleagues explain, these units “represent the first-order network’s error not as a training signal but as a distributed activation pattern, which the second-order network can then access by using a weighted sum of these signed errors to decide on whether to place a high or a low wager.” (2010, p. 196).[2]

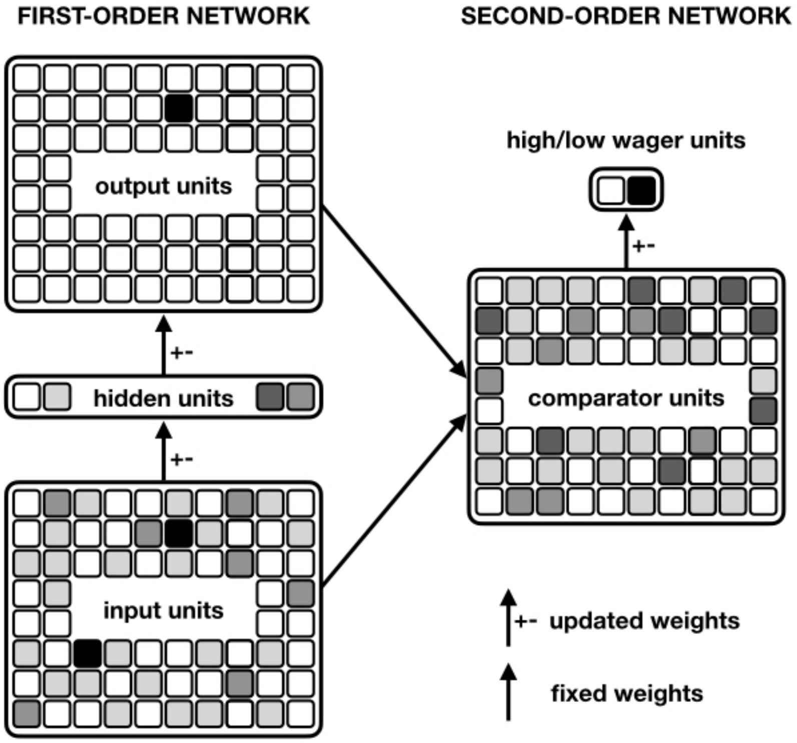

Figure 2: An illustration of a MN used for simulation of the perceptual discrimination tasks. The first-order network has one hidden layer connected to the input and output layers with weights that are updated through backpropagation in training. The hidden layer of the second-order network is also connected to the first-order input and output layers, but its weights are fixed before the first-order system is trained. The connections between comparator units and the wagering output layer of the second-order network are optimized through feedback, but this also happens separately from the first-order training process. Adapted from Pasquali et al. (2010).

[2] From a technical point of view, the function of the second-order network is very similar (if not outright identical) to a type 2 signal detection task. Although there are several competing measures of this type of performance, all of them aim to quantify the degree to which observers’ confidence ratings track their performance on the first-oder task (Maniscalo & Lau, 2011). However, unlike many of type 2 signal detection measures, MNs “ second-order classification does not depend on the same signal as the first-order task. Instead of wagering high or low based on signal strength, the second-order network re-represents the first-order error, thus basing itself more on a consequence of signal coherence.” (Pasquali et al., 2010, p. 188).

However, MN networks earn their name not only in virtue of the hierarchical structure of their representations, but also thanks to the functional relationship between the higher-order network's representations and the lower-order network's performance on its task. Firstly, the comparison patterns (i.e., the weights connecting it to the first-order input and output layers) embedded in that layer are learned automatically and without feedback. Secondly, the weights connecting the units in the higher-order comparison layer to the two wagering output units are trained in a pre-training phase which is independent of the first-order network's training and testing. As Pasquali et al. (2010) explain, it is these "two properties allow for the second-order network to access the relevant first-order knowledge in a manner that is independent of the causal chain in which that knowledge is embedded." (p. 184). In other words, the higher-order network does not simply "match specific first-order inputs and outputs to a high or a low wager" (ibid.) but learns to re-represent the state of the lower-order network in such a way that it can be described as having the 'knowledge' about whether the lower-order network will successfully detect the presence or absence of a stimulus on a given trial. Therefore, the key feature that separates MN models from most other neural networks is that they encode information in meta-representations that render the first-order content useful *for* the system, rather than being merely stored *in* the system. (Clark & Karmiloff-Smith, 1992).[3]

In short, what makes MNs an attractive model of introspection is not just that they closely match human performance on both the first- and second-order tasks (Pasquali et al., 2010) or utilize meta-representations that target lower-order states without copying their content. Rather, it is that the higher-order states responsible for the network's second-order task performance come to represent the *subjective* reliability of the lower-

[3] It should be noted that the network described here is not the only kind of MN possible. For example, Pasquali et al. (2010) discuss a second kind of architecture in which the higher-order network has direct access to the internal representations of the first-order network's hidden layer. Although this kind of MN network might provide yet another attractive formalization of introspection, I've decided not to discuss such networks for two reasons. Pasquali and colleagues have only applied such networks to the Iowa Gambling task, which is significantly different from the perceptual cases discussed in the main body of this article.Furthermore, the higher-order states of such networks are not independent from the feedback signal on the first-order task, raising questions whether the higher-order network forms genuine meta-representations about the reliability of the first-order network or whether it merely learns to solve its task by relying on the first-order feedback.

order network's performance. In other words, an MN trained to perform PDW in a perceptual paradigm "knows that it knows" about a stimulus, "as would be the case for knowledge held consciously by a human agent." (Pasquali et al. 2010, p. 183). It is this kind of knowledge that is predominantly being investigated by researchers studying metacognition in humans (see e.g., Maniscalco & Lau, 2012).

3. Evaluating models against the introspective devices framework

The two models of introspection presented above are highly idealized and certainly do not capture the full complexity of the target phenomenon or reflect the range of possible introspective devices. However, the simplified nature of these models is an advantage in the present context, as it should facilitate evaluating them from the perspective of the introspective devices framework. In what follows, I will assess both models according to the three main dimensions proposed by K&F.

3.1. The directness dimension

The first of K&F's dimension intends to capture "how close and direct the informational relation will be between the introspective states and the mental states they represent." (K&F, p. 9). In other words, this dimension presents a graded measure of separation or mediation between introspection and its targets.

Starting with MNs, it seems fairly easy to locate the model proposed by Pasquali and colleagues on the scale measuring introspection's relation to its target states. Recall that, in those architectures, the activations of the units in the higher-order network's hidden layer represent a mismatch between the lower-order network's inputs and outputs, and that this is achieved without access to any feedback or error signal that the first-order network receives. This is why the authors claim that the function of the second-order network as a whole "effectively comes down to setting a decision criterion on the first-order network's error distribution" (Pasquali et al., 2010, p. 184). Therefore, even though the higher-order network is directly connected to the lower-order one, the wagering decision layer has only indirect access to the states of the first-order network.[4]

Things look more complicated in the case of iSDT. Morales introduces his view as a version of the 'inner sense' theories (see e.g., Armstrong, 1968, and Lycan, 1996), according to which introspection is modeled after perception. This is an intuitive proposal, given that SDT is commonly applied to perception. So, since Inner sense theories are categorized as offering a direct view of introspection on K&F's framework, iSDT should be located close to those theories? Unfortunately, the issue is not that simple.

Morales himself is quick to point out that the analogy with perception should be taken loosely, as introspection is no more perception "than perception is receiving radio signals" (forthcoming, p.9). What complicates matters further is that introspection under iSDT, just like perception under SDT, can be broken down into two separate components — the internal introspective response (i.e., experience or signal) and the introspective judgement (criterion setting in SDT). iSDT's placement along the directness-indirectness spectrum will depend on which of the two components we take to be more important. Focusing on the introspective response would categorize iSDT as a direct theory of introspection, since one makes decisions about an experience based on its mental strength, hence the experience itself seems to be involved in the process of introspection. However, focusing on the introspective judgement will lead to the opposite classification, as the process involved in the setting of the introspective response criterion is widely considered to be inferential in nature and more akin to a higher-decision process employed in MNs than any form of direct access (Lau, 2008).

[4] As has been mentioned in a previous footnote, the network discussed here does not exhaust the space of possible MN architectures. Models with direct connections between the hidden states of the lower- and higher-order networks are entirely feasible. It should be noted, however, that presence of such connections does not guarantee direct access in the sense employed by K&F, since different computational implementations may rely on additional assumptions that will complicate the picture. Similarly, it should not be assumed that MNs with more hidden layers in the second-order network will automatically trade in more indirect representations, as meta-representations could be distributed across several layers. Hence, each MN architecture should be evaluated on its own merit.

3.2. The conceptual dimension

K&F's second dimension concerns the distinction between conceptual and non-conceptual forms of introspective outputs or whether introspection "will generate representational states with a format that is conceptual, akin to beliefs or propositionally structured perceptions, or non-conceptual, akin to sensations." (K&F, p. 10).

Given what has been said so far, it should come as no surprise that both iSDT and MNs offer a picture of introspection that is conceptual in nature. Again, matters seem simpler in the latter case, as MNs are explicitly designed to simulate the PDW task and produce outputs that correspond to a binary choice of placing predefined bets. While bets are not propositional attitudes, in the context of the wagering task, they are used as an implicit measure of subjective credence about one's discrimination decision, and credences of this type seem much closer to traditionally construed beliefs than perceptual representations. Critics of connectionism might baulk at this by arguing that artificial networks do not hold explicit beliefs, but this would be mistaking the map for the territory, since MNs are highly idealized models that are supposed to provide a computational sketch of how meta-knowledge could be acquired and are not intended as realistic depictions of full-fledged artificial agents. Nevertheless, it is important to note that MNs could easily be modified to work within a larger space of possible responses (e.g., by increasing the size of the second-order network's output layer) and are, in principle, applicable to a wider range of tasks.

Returning to iSDT in the present context reveals the source of the indeterminacy uncovered in the previous subsection. iSDT formalizes introspection as a decision process in which inputs are classified according to a chosen criterion. The output of such process boils down to the categorization of a certain introspective response *i* with the mental strength *x* as a mental experience of class *C1* or *C2*. Thus, the output of introspection on iSDT is something akin to a judgment with the propositional content "*i* is *Cx*". Thus, iSDT produces conceptual states as its outputs. Or does it?

K&F distinguish between *discriminating* mental states and *conceptualizing* them. As they point out, introspection "might distinguish two types of mental state without characterizing them in any substantive way; it might simply represent them as this type and that type." (K&F, 2023, p. 16). Does iSDT discriminate or conceptually characterize target mental states? This is a complex question that requires a careful investigation in a separate treatment. Here, I would like to merely highlight the tension between the ways in which K&F's characterization of the conceptual dimension and their account of possible introspective repertoires (i.e., discrimination vs characterization) classify iSDT by the outputs it produces.

3.3 The flexibility dimension

The third dimension along which K&F propose to locate different forms of introspection concerns the degree to which the functioning of an introspective device can be modified by the system in which it is embedded. As the authors explain: "a cognitive system might be able to control when introspection occurs and where it is directed (say, whether to beliefs or to perceptions) but unable to control how it operates (what processes it uses or what format its outputs take)" (K&F, p. 11). For present purposes I will assume that this dimension is an aggregate measure of two related features — the degree of freedom about *where* in the cognitive system or at *what* mental state introspection can be deployed as well as the number of degrees of freedom that are available to the system regarding *how* introspection can be modified.

Yet again, MNs present a rigid model of introspection that seems to be fairly constrained regarding where and how it can be deployed. The second-order network monitors the first-order network's input and output layers with fixed connections, hence there is no possibility for re-deployment elsewhere. The ways in which introspection can be modified are similarly restricted, since the weights on the connections between the two networks are fixed and only the weights between the second-order network's hidden and output layers are modifiable in training. MNs, therefore, offer a picture of introspection that is very domain specific. However, it is worth noting that this is, once again, a limitation imposed by the idealized nature of these models. MNs have been successfully applied (with or without modification) to domains outside of perception like artificial grammar learning and the Iowa gambling task (Pasquali et al., 2010), the main limitation here is the need to re-train the network when switching tasks. However, it is quite possible that large enough MNs could display some level of domain generality and flexibility.

Contrary to MNs, iSDT shows a rather large degree of flexibility regarding where and how it can be deployed. Morales assumes that introspection can not only be aimed at any conscious experience but can also be a source of subjective illusions and false reports in cases where internal noise is confused for genuine internal signal. Thus, introspection can not only be applied to conscious mental states but can also create them. Similarly, iSDT has at least one parameter — the placement of the decision criterion *c* — that seems to be under the subject's control. Yet, as in the case of MNs, it is worth stressing that the flexibility of this account of introspection seems to be largely due to the generality of the underlying SDT model.

4. Lessons for a more introspective future

The aim of the present article was to see whether the introspective devices framework can be applied to existing models of introspection coming from outside of philosophy, and whether such applications will deliver novel insights about those models or reveal the framework's limitations.

Considering what has been said in the previous section of this paper, it seems fairly clear that the framework can, in fact, be successfully applied to models that have not been considered by its authors. Furthermore, it is highly likely that locating different models of introspection within K&F's conceptual structure will offer a novel perspective on their commitments and could be used to guide the further development of said models as well as the expansion and refinement of the framework itself. Below I will present what I take to be some of the major takeaways from the exercise carried out in this paper.

Starting with lessons that the framework can offer to the modelers, discussions on the format of introspective outputs and introspection's relation to its inputs have both revealed an ambiguity in how iSDT is to be interpreted. Proponents of that view could claim, for example, that introspection is either direct or indirect (or a combination of both) while maintaining that it delivers conceptual representations. It seems that, the kinds of introspective devices that proponents of iSDT may end up searching for will depend not just on the measurable parameters of the model, but also on which interpretation of iSDT is adopted. In other words, there are multiple devices that could serve as potential realizers of iSDT. Thus, applying K&F's framework to iSDT reveals that the model under-constrains the space of possible introspective devices it is compatible with. This is not an outright flaw, since no one knows which account of introspection is the correct one, but it carries a valuable lesson about the many-to-one relation between theories of introspection and formal models.

A further lesson emerging form analyzing models of introspection through the lens of the framework is that this kind of scrutiny can offer novel directions in which models can be developed or expanded. For example, the analysis of MNs invites an investigation into whether such networks are bound to produce discrete outputs and whether it might be possible to design MNs that output credences in the form of confidence intervals rather than sets of predefined binary variables. A further open question is whether such new kinds of MN outputs could be considered as non-conceptual rather than conceptual in nature.

The main lessons about the structure of K&F's framework that can be extracted from present analysis is that models will tend to occupy fuzzily individuated regions of the conceptual space of introspective devices rather than correspond to well-defined vectors. The two main factors responsible for this are: the aforementioned one-to-many model-device mapping as well as possible problems related to how models are to be located along one or more of K&F’s chosen dimensions, as revealed during the discussion of iSDT’s conceptual commitments (see Figure 3).

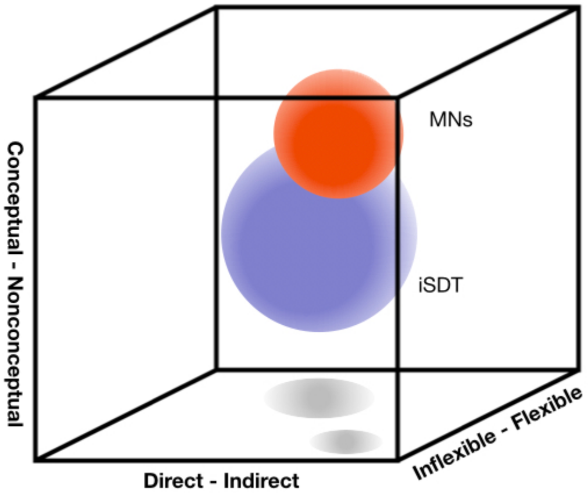

Figure 3: A simplified rendition of K&F’s space of possible introspective devices. MNs and iSDT models are represented as differently sized colored spheres occupying separate portions of the space to indicate that formal models of introspection can be compatible with multiple kinds of introspective devices. Adapted from K&F (2023).

Fortunately, neither of the above problems threatens the framework as such. Indeed, both issues mentioned above can be taken as indicative of the limitations inherent in categorizing complex theoretical hypotheses along only three broad dimensions. One possible way for reducing ambiguity and further refining the framework is to expand the dimensionality of its conceptual space, thereby increasing the specificity of how possible introspective devices are categorized. Introspective reliability is one of the widely discussed dimensions that could be easily added to K&F’s framework. Precision (or the degree to which introspection transforms or distorts its targets) and phenomenal opacity (or the degree to which the subjects are aware they are deploying their introspective faculties) are examples of other possible dimensions. Finally, it is worth noting that K&F also consider differences in the *introspective repertoires* which can vary even between accounts that co-occupy the same region in their conceptual space (p. 17). This part of K&F’s story offers yet another avenue for future expansion and enhancement. Importantly, K&F are not only open to, but actively encourage this kind of expansion of their framework, noting that their “diagram is merely a first, tentative step” (p. 15) on the road to a fully mature science of introspection.

Acknowledgements:
The author would like to thank François Kammerer, Keith Frankish, and two reviewers for helpful comments and an invitation to write this commentary.

Funding information:
Funded by the European Union (ERC, ‘EXPERIENCE’, #101055060). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References:

- Armstrong, D. M. (1968) *A Materialist Theory of the Mind*, London: Routledge & Kegan Paul.
- Clark, A., & Karmiloff-Smith, A. (1993) The cognizer’s innards: a psychological and philosophical perspective on the development of thought. *Mind and Language*, 8, pp. 488–519.
- Green, D.M., & Swets J.A. (1966) *Signal Detection Theory and Psychophysics*, New York: Wiley.
- Kammerer, F., & Frankish, K. (2023) What forms could introspective systems take? A research programme, *Journal of Consciousness Studies*, **???**.
- Lau, H.C. (2008) A higher order Bayesian decision theory of consciousness, *Progress in Brain Research*, 168, pp. 35-48
- Lycan, W. G. (1996) *Consciousness and Experience*, Cambridge, MA: MIT Press.
- Maniscalco, B., & Lau, H.C. (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness & Cognition*, 21(1), pp. 422-30.

- Morales, J. (forthcoming) Introspection Is Signal Detection, *British Journal for the Philosophy of Science*, https://dei.org/10.1086/715184.
- Nieder, A., Wagener, L., & Rinnert, P. (2020) A neural correlate of sensory consciousness in a corvid bird, *Science* 369, pp. 1626-1629.
- Pasquali, A., Timmermans, B., & Cleeremans, A. (2010) Know thyself: Metacognitive networks and measures of consciousness, *Cognition*, **117**(2), pp.182-190.
- Persaud, N., McLeod, P., & Cowey, A. (2007) Post-decision wagering objectively measures awareness, *Nature Neuroscience*, 10, pp. 257–261.
- Tanner, W.P., & Swets, J.A. (1954) A decision-making theory of visual detection, *Psychological Review*. **61** (6), pp. 401–409.
- Timmermans, B., & Cleeremans, A. (2015) How can we measure awareness? An overview of current methods. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research*, Oxford: Oxford University Press, pp. 21-46.