Preprint of November 20, 2015. Forthcoming in *Thought: A Journal of Philosophy*.

On Being in an Undiscoverable Position

Wesley H. Holliday

University of California, Berkeley

The Paradox of the Surprise Examination has been a testing ground for a variety of frameworks in formal epistemology, from epistemic logic to probability theory to game theory and more (see Chow 2011 for references). In this paper, I will treat a related paradox, the Paradox of the Undiscoverable Position (from Sorensen 1982, 1988), as a test case for the possible-worlds style representation of epistemic states. I will argue that the paradox can be solved in this framework, further illustrating the power of possible-worlds style modeling. The solution will also illustrate an important distinction between anti-performatory and unassimilable announcements of information.

The Paradox of the Undiscoverable Position is based on a game played in the following figure:

1	2	3
4	5	6
7	8	9

Sorensen (1988) describes the game as follows:

> The object of the game is to discover where you have been initially placed. The seeker may only move Up, Down, Left, or Right, one box at a time. The outer edges are called walls. If the seeker bumps into a wall, say by moving left from 1,

> his move is recorded as L* and his position is unchanged. Bumps help the seeker discover his initial position. For instance, if he is at 7 and moves U, U, L*, the seeker can deduce that he must have started from 7. The seeker has discovered where he started from if he obtains a completely disambiguating sequence of moves, i.e. a sequence which determines the seeker's initial position. (320)

Sorensen describes the paradox as follows:

> If the seeker is given only two moves, it is possible to put him in an undiscoverable position. For instance, if he is put in position 4, every possible two move sequence is compatible with him having started from some other position. Now suppose the seeker is told 'You have been put in an undiscoverable position'. He disagrees and offers the following *reductio ad absurdum*:[1] Suppose I am in an undiscoverable position. [1] It follows that I cannot be in any of the corners since each has a completely disambiguating sequence. For instance, if I am in 3, I might move U*, R*, and thereby deduce my position. [2] Having eliminated the corners, I can also eliminate 2, 4, 6, and 8, since any bumps resulting from a first move completely disambiguates. For instance, U* is sufficient to show that I am in 2. Since only 5 remains, I have discovered my position. [3] The absurdity of the supposition is made further manifest by the existence of eight other arguments with eight distinct conclusions as to my initial position. For example, I could conclude that I am in 6 by first eliminating the corners, then 2, 4, 8, and then 5 (by sequence L, L*, leaving only 6 remaining). (320-21)

[1] The numbers in square brackets are my additions.

Where is the mistake in the seeker's reasoning toward his absurd conclusion?

Call an element of {1, 2, 3, 4, 5, 6, 7, 8, 9} a *position*. Call a sequence of elements from {L, R, U, D, L*, R*, U*, D*} a *sequence of moves*. Say that a sequence S of moves is *executable* from a position P if and only if it is possible, starting from P, to make moves (including bumps) in the order that S indicates. Say that for a position P and a set X of positions, P is *definable in X using two moves* if and only if there is a sequence of two moves that is executable from P but not executable from any other P' in X. Since we are not interested in other numbers of moves, in what follows let us simply say 'definable', dropping the qualifier 'using two moves'.

The solution to the paradox begins by disambiguating the announcement. There are at least four things the announcer could mean by 'you have been put in an undiscoverable position':

Undiscoverable$_1$: you are in a position P that is not definable in {1, 2, 3, 4, 5, 6, 7, 8, 9}.

Undiscoverable$_2$: where E_{current} is the set of positions compatible with your current knowledge, you are in a position P that is not definable in E_{current}.

Undiscoverable$_3$: where E_{after} is the set of positions compatible with your knowledge right after this announcement, you are in a position P that is not definable in E_{after}.

Undiscoverable$_4$: for any time t, where E_t is the set of positions compatible with your knowledge at t, you are in a position P that is not definable in E_t.

Below I will argue that for each of these ways of understanding the announcement, the paradoxical reasoning rehearsed by Sorensen is incorrect reasoning. In each case, one of Sorensen's steps [1], [2], and [3] does not go through.

In the spirit of possible-worlds modeling, I will call the set of positions compatible with the seeker's knowledge at a given time the seeker's *epistemic state* at that time. Thus, at the start of the game, before any announcement, his epistemic state is $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Whatever position P the seeker is in, P is an element of his epistemic state, because the truth is always compatible with one's knowledge. Since we assume the seeker is in some position, it follows that his epistemic state can never be $\varnothing$.

First, suppose that the announcer announces **Undiscoverable$_1$**, and the seeker updates his epistemic state accordingly. One can easily check that

$$\{P \mid P \text{ satisfies } \textbf{Undiscoverable}_1\} = \{2, 4, 5, 6, 8\},$$

so $\{2, 4, 5, 6, 8\}$ is the seeker's new epistemic state. In more detail: the seeker eliminates the corners 1, 3, 7, and 9, as in Sorensen's step [1], because each of these is definable in $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. However, if the announcement is understood as **Undiscoverable$_1$**, then Sorensen's step [2] does not go through: being in each of 2, 4, 6, and 8 is consistent with **Undiscoverable$_1$**, because each of 2, 4, 6, and 8 is not definable in $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Of course, 2, 4, 6, and 8 are each definable in $\{2, 4, 5, 6, 8\}$. But that point is irrelevant, since **Undiscoverable$_1$** refers to the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, not to the set $\{2, 4, 5, 6, 8\}$. Finally, note that if the announcer were to announce **Undiscoverable$_1$** a *second* time, this would be a true announcement (assuming the first announcement was true), but it would offer the seeker no new information about his position. His epistemic state is stuck at $\{2, 4, 5, 6, 8\}$.

Second, suppose that instead of announcing **Undiscoverable$_1$**, the announcer announces **Undiscoverable$_2$**, and the seeker updates his epistemic state accordingly. Remember that before the announcement, the set of positions compatible with his knowledge is {1, 2, 3, 4, 5, 6, 7, 8, 9}. Thus, the *indexical* phrase 'set of positions compatible with your current knowledge' at the beginning of the announcement refers to {1, 2, 3, 4, 5, 6, 7, 8, 9}. Just as before, we have

{P | P satisfies **Undiscoverable$_2$** with E_{current} = {1, 2, 3, 4, 5, 6, 7, 8, 9}} = {2, 4, 5, 6, 8},

so {2, 4, 5, 6, 8} is the seeker's new epistemic state after the announcement of **Undiscoverable$_2$**; as before, Sorensen's step [2] does not go through.

But now suppose that **Undiscoverable$_2$** is announced to the seeker a *second* time, and suppose that this second announcement is true. Since the seeker's epistemic state after the first announcement is {2, 4, 5, 6, 8}, the indexical phrase 'set of positions compatible with your current knowledge' at the beginning of the *second* announcement refers to {2, 4, 5, 6, 8}. Now one can check that

{P | P satisfies **Undiscoverable$_2$** with E_{current} = {2, 4, 5, 6, 8}} = {5},

so the seeker can determine his position. But there is nothing absurd about this result, because repeated announcement of **Undiscoverable$_2$** cannot produce a nonempty epistemic state disjoint from {5}, so Sorensen's step [3] does not go through.

To underscore the last point, suppose that **Undiscoverable$_2$** is announced to the seeker a *third* time. Since the seeker's epistemic state after the second announcement is {5}, the indexical phrase 'set of positions compatible with your current knowledge' at the beginning of the *third* announcement refers to {5}. Then since

{P | P satisfies **Undiscoverable$_2$** with E_{current} = {5}} = ∅, the third announcement of **Undiscoverable$_2$** is *false*, because the seeker is in position 5 in this case. Thus, the *second* announcement of **Undiscoverable$_2$** was what Hintikka (1962) calls an *anti-performatory* announcement: "If you know that I am well informed and if I address the words . . . to you, these words have a curious effect which may perhaps be called anti-performatory. You may come to know that what I say *was* true, but saying it in so many words has the effect of making what is being said false" (68-69). Or to put the point more carefully: it has the effect that a subsequent announcement using the same words would be a false announcement.[2] Even if there is no third *announcement* of **Undiscoverable$_2$**, when the seeker thinks about **Undiscoverable$_2$** at a time when $E_{\text{current}} = \{5\}$ (after the second announcement), **Undiscoverable$_2$** will then be false.

Third, suppose that instead of announcing **Undiscoverable$_1$** or **Undiscoverable$_2$**, the announcer announces **Undiscoverable$_3$**, and the seeker updates his epistemic state accordingly. Unlike in the previous cases, it is not so clear what the seeker's updated epistemic state is as a result of the announcement of **Undiscoverable$_3$**. The announcement of **Undiscoverable$_3$** indicates that the seeker's position is in the set

$$\{P \mid P \text{ is not definable in } E_{\text{after}}\},$$

where E_{after} is his epistemic state resulting from that very announcement. So if in the

[2] That repeated announcement of a sentence containing indexical expressions referring to epistemic states may start with a true announcement and wind up in a false announcement is a phenomenon that has been thoroughly investigated in the field of *dynamic epistemic logic* (see van Benthem 2004, van Ditmarsch and Kooi 2006, Holliday and Icard 2010, and Holliday, Hoshi, and Icard 2013). Gerbrandy (2007) argues that this phenomenon is behind the surprise exam paradox. I argue elsewhere (Holliday 2015) that it is the phenomenon of *unassimilable* announcements described below, not the phenomenon of *anti-performatory* announcements, that is relevant to the surprise exam paradox.

resulting epistemic state the seeker knows what the announcement of **Undiscoverable$_3$** indicates about his position, then all of the positions compatible with his knowledge will belong to the set $\{P \mid P \text{ is not definable in } E_{\text{after}}\}$, which is to say:

$$E_{\text{after}} \subseteq \{P \mid P \text{ is not definable in } E_{\text{after}}\}.$$

But one can check that for every nonempty set X of positions, at least one of the positions P in X is definable in X. So the only value for E_{after} that satisfies the inclusion statement above is $\varnothing$, and as before, $\varnothing$ cannot be the seeker's epistemic state. Thus, although the seeker will have *some* epistemic state E_{after} after the announcement of **Undiscoverable$_3$** (more on this below), it cannot satisfy the inclusion statement above. In other words, it cannot be that in his updated epistemic state, the seeker knows what the announcement of **Undiscoverable$_3$** indicates about his position. In light of this fact, it seems natural to call the announcement of **Undiscoverable$_3$** *unassimilable* for the seeker.

Let us make this notion of unassimilability more precise. In the context of Sorensen's game, let us say that the *proposition* expressed by an announcement is the set of positions P such that it is compatible with the truth of the announcement that the seeker is in P. As philosophers of language have long discussed, which proposition is expressed by a token utterance may depend on factors that vary across contexts, such as the values associated with indexicals, so distinct token utterances of the same sentence may express distinct propositions.[3] (I am using 'announcement' and 'utterance' interchangeably.) In the case of interest to us, the proposition expressed by an announcement of **Undiscoverable$_3$** depends on the value of 'E_{after}'. Fixing a value for

[3] The following analysis of **Undiscoverable$_3$** would fit nicely into the picture of "incremental contents" in §5.4 of Perry 2011, but we will not go into the details here.

‘E_{after}’, the proposition expressed is $\{P \mid P$ is not definable in $E_{\text{after}}\}$. The announcement of **Undiscoverable$_3$** is *unassimilable* in the sense that there is no possible value for ‘E_{after}’ such that given that value, the proposition expressed by **Undiscoverable$_3$** is *known* in the epistemic state denoted by ‘E_{after}’, i.e., such that $E_{\text{after}} \subseteq \{P \mid P$ is not definable in $E_{\text{after}}\}$, as explained above. In general, let us say that a sentence is unassimilable for the seeker if and only if for every epistemic state E, if E is the seeker’s new epistemic state as a result of an announcement of that sentence, then the proposition expressed by that announcement is not known in E. An announcement is unassimilable for the seeker if and only if it is an announcement of a sentence that is unassimilable for the seeker.

Although the announcement of **Undiscoverable$_3$** is unassimilable, it is noteworthy that this announcement may be *true*. For example, if the seeker is in position 4 and $E_{\text{after}} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, then it is true that the seeker is in a position P that is not definable in E_{after}. Moreover, since a *second* announcement of **Undiscoverable$_3$** would also be unassimilable for the seeker, it may also be true. Thus, the first announcement of **Undiscoverable$_3$** is not necessarily *anti-performatory* in the way that a second announcement of **Undiscoverable$_2$** would be.[4]

[4] Are any unassimilable announcements anti-performatory? It depends on the precise definition of ‘anti-performatory’. Hintikka’s characterization suggests that with an anti-performatory announcement, (a) the hearer may come to know that the proposition that was expressed by the announcement is true, but (b) as a result of that true announcement and the hearer’s resulting epistemic update, a subsequent token announcement of the same type would express a proposition that is false. This can happen when I tell you, “You don’t know it, but I have a dime in my pocket.” By contrast, in the case of an unassimilable announcement, the hearer cannot even come to know as a result of the announcement that the proposition that was expressed by the announcement is true. So an announcement of “You don’t know it, but I have a dime in my pocket” is anti-performatory but not unassimilable. To decide whether there are any

If the announcement of **Undiscoverable$_3$** is *true*, then this constrains the possible values of P and E_{after}: P must be in E_{after} and not definable in E_{after}. As one can check, this implies that one of the following holds:

$P = 2$ and $\{1, 2, 3\} \subseteq E_{\text{after}}$ and $[5 \in E_{\text{after}}$ or $8 \in E_{\text{after}}]$;

$P = 4$ and $\{1, 4, 7\} \subseteq E_{\text{after}}$ and $[5 \in E_{\text{after}}$ or $6 \in E_{\text{after}}]$;

$P = 6$ and $\{3, 6, 9\} \subseteq E_{\text{after}}$ and $[5 \in E_{\text{after}}$ or $4 \in E_{\text{after}}]$;

$P = 8$ and $\{7, 8, 9\} \subseteq E_{\text{after}}$ and $[5 \in E_{\text{after}}$ or $2 \in E_{\text{after}}]$;

$P = 5$ and $5 \in E_{\text{after}}$ and $[4 \in E_{\text{after}}$ or $1 \in E_{\text{after}}$ or $2 \in E_{\text{after}}]$ and
$[2 \in E_{\text{after}}$ or $3 \in E_{\text{after}}$ or $6 \in E_{\text{after}}]$ and
$[6 \in E_{\text{after}}$ or $9 \in E_{\text{after}}$ or $8 \in E_{\text{after}}]$ and
$[8 \in E_{\text{after}}$ or $7 \in E_{\text{after}}$ or $4 \in E_{\text{after}}]$ and
$[2 \in E_{\text{after}}$ or $8 \in E_{\text{after}}]$ and $[4 \in E_{\text{after}}$ or $6 \in E_{\text{after}}]$.

Now it suffices to observe that no matter what E_{after} is, the seeker does not eliminate all of 2, 4, 6, and 8, so Sorensen's step [2] does not go through.

Let us digress to consider the most plausible value for E_{after} after the announcement of **Undiscoverable$_3$**. Although P and E_{after} are not uniquely determined by just the truth of the announcement of **Undiscoverable$_3$**, one could argue that since the

announcements that are both anti-performatory and unassimilable, we need to decide whether to count (a) above as necessary for anti-performativeness. If (a) is necessary, then unassimilable announcements are never anti-performatory. If (a) is not necessary, then the question is whether it can happen that as a result of a true but unassimilable announcement and the hearer's resulting epistemic update, a subsequent token announcement of the same type would express a proposition that is false. This is not obvious, because it is not obvious what the hearer's resulting epistemic state *is* in the case of an unassimilable announcement, as discussed below. For our purposes in this paper, we need not settle the question. (Thanks to an anonymous referee for pressing the issue of the logical relations between the notions of *unassimilable* and *anti-performatory*.)

announcement of **Undiscoverable₃** was unassimilable by the seeker, if the only thing that happened since the start of the game was the announcement of **Undiscoverable₃**, then E_{after} should be $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Based on the enumeration of cases above, *we* as bystanders to the game were able to deduce, assuming the announcement of **Undiscoverable₃** to the seeker was true, that the seeker's position is in $\{2, 4, 5, 6, 8\}$. One might think it follows that the seeker also knows this, so $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$. But it does not follow. If the seeker's position is in $\{2, 4, 6, 8\}$, then it cannot be that the announcement was *true* and that $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$. It is only if the seeker's position is 5 that the truth of the announcement is *consistent* with $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$; and mere consistency does not show that the announcement would in fact have the effect that $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$ when $P = 5$. Since a true announcement of **Undiscoverable₃** cannot result in $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$ when $P = 4$, why should a true announcement of **Undiscoverable₃** result in $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$ when $P = 5$? Whether $P = 4$ or $P = 5$, the seeker's initial epistemic state is the same; he hears the same words announced; and we can assume that the announcer only makes true announcements. If we adopt the principle that the seeker's epistemic state after the announcement of **Undiscoverable₃** should be invariant under changing the seeker's initial position within $\{2, 4, 5, 6, 8\}$, then this uniquely determines $E_{\text{after}} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, because $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ is the only value for E_{after} that is permitted by every one of the cases enumerated above.[5] In this case, the invariance principle implies that not even Sorensen's step [1] goes through.

[5] Suppose that something else happens in addition to the announcement of **Undiscoverable₃**, with the result that $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$. Further suppose that *we* know from the announcement that the seeker's position is in $\{P \mid P \text{ is not definable in } E_{\text{after}}\}$, and we somehow also know of his epistemic state that $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$. Then we can deduce that the seeker's position is 5,

Finally, suppose that instead of announcing **Undiscoverable$_1$**, **Undiscoverable$_2$**, or **Undiscoverable$_3$**, the announcer announces **Undiscoverable$_4$.** Observe that a true announcement of **Undiscoverable$_4$** implies the following for the seeker, by universal instantiation: where E_{after} is the set of positions compatible with your knowledge right after this announcement of **Undiscoverable$_4$**, you are in a position P that is not definable in E_{after}. Thus,

$$\{P \mid P \text{ satisfies } \textbf{Undiscoverable}_4\} \subseteq \{P \mid P \text{ is not definable in } E_{\text{after}}\},$$

which implies

$$E_{\text{after}} \nsubseteq \{P \mid P \text{ satisfies } \textbf{Undiscoverable}_4\},$$

for otherwise we would have

$$E_{\text{after}} \subseteq \{P \mid P \text{ is not definable in } E_{\text{after}}\},$$

which we have seen is not possible. Given the non-inclusion stated above, the seeker does not come to know what the announcement of **Undiscoverable$_4$** indicates about his position. Like the announcement of **Undiscoverable$_3$**, the announcement of **Undiscoverable$_4$** is unassimilable for the seeker. Additional observations, similar to those made about **Undiscoverable$_3$**, could be made about **Undiscoverable$_4$**, but the conclusion is this: the announcement of **Undiscoverable$_4$** does not allow the seeker to eliminate all of 2, 4, 6, and 8, so Sorensen's step [2] does not go through.

since this is the only case in which the truth of the announcement is consistent with $E_{\text{after}} \subseteq \{2, 4, 5, 6, 8\}$. Similarly, if the seeker himself somehow came to know at a later time that his initial position was in $\{P \mid P \text{ is not definable in } E_{\text{after}}\}$ (which we have seen he cannot come to know as a result of the announcement), and he somehow also came to know that his epistemic state after the announcement was a subset of $\{2, 4, 5, 6, 8\}$, then he could deduce that his position was 5. But there is nothing absurd about this result, because he could not deduce in a similar fashion that he was in a different position, so Sorensen's step [3] does not go through.

For all of the ways we have thought of understanding the announcement that 'you have been put in an undiscoverable position', one of Sorensen's steps does not go through. Thus, one cannot reach the absurd conclusion in Sorensen's passage. If some philosophers still think there is a paradox here, then they must tell us how to understand the announcement to generate a paradox. Otherwise it seems that the paradox is solved. The distinction between anti-performatory and unassimilable announcements drawn above also plays a role in solving the surprise exam paradox, but that is a longer story (Holliday 2015).

There is also a methodological reminder in the analysis of this paper: despite the grandiose connotation of the term 'possible-worlds modeling', when modeling epistemic states in this style, it often suffices to use objects much more modest than *complete worlds* to represent a certain aspect of an agent's knowledge or ignorance. Here our "possible worlds" were possible positions in a game. Kripke's (1972, 16) "(miniature) 'possible worlds'" included possible rolls of dice. Whatever one makes of the idea of possible *worlds*, there is no denying the utility of modeling with possibilities.

Acknowledgements

I wish to thank Seth Yalcin and an anonymous referee for helpful comments.

References

van Benthem, Johan. "What One May Come to Know." *Analysis* 64.2 (2004): 95-105.

van Ditmarsch, Hans and Barteld Kooi. "The Secret of My Success." *Synthese* 151.2 (2006): 201-232.

Chow, Timothy Y. "The Surprise Examination or Unexpected Hanging Paradox." arXiv:math/9903160v4 [math.LO] (2011). Previously published in *American Mathematical Monthly* 105 (1998): 41-51.

Gerbrandy, J. "The Surprise Examination in Dynamic Epistemic Logic." *Synthese* 155.1 (2007): 21-33.

Hintikka, Jaakko. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. London: College Publications, 2005. Reprint of the 1962 Cornell University Press edition.

Holliday, Wesley H. and Thomas F. Icard, III. "Moorean Phenomena in Epistemic Logic," in *Advances in Modal Logic* 8, edited by L. Beklemishev, V. Goranko, and V. Shehtman. London: College Publications, 2010, 178-199.

Holliday, Wesley H., Tomohiro Hoshi, and Thomas F. Icard, III. "Information Dynamics and Uniform Substitution." *Synthese* 190 (2013): 31-55.

Holliday, Wesley H. "Simplifying the Surprise Exam." Manuscript, 2015.

Kripke, Saul A. 1972. *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Perry, John. *Reference and Reflexivity*. 2nd Edition. Stanford, CA: CSLI Publications, 2011.

Sorensen, Roy A. "Recalcitrant Variations of the Prediction Paradox." *Australasian Journal of Philosophy* 69.4 (1982): 355-362.

Sorensen, Roy A. *Blindspots*. Oxford: Clarendon Press, 1988.