The Vain Mathematicians

> A university has ten vain mathematicians. Each is so egocentric that if he ever learns of a mistake in his work, he resigns the next Friday. (Note [1] contains an apology for our use of pronouns.)
>
> Resignations increase the teaching workload, so the mathematicians are discreet when they find mistakes in their colleagues' work. Anyone who detects a mistake tells everyone *except* the mistake-maker. Each mathematician has made at least one mistake, and every mistake is known to everyone else, so each mathematician thinks himself the only one who is error-free.
>
> One Monday a super-mathematician, known to be infallible, comes to visit. She looks at everyone's work, assembles the whole group, and says to them all "Someone here has made a mistake." Assuming that mathematicians are perfect reasoners, what happens?

This puzzle is a chestnut with many isomorphic variants. Sometimes the mathematicians commit suicide rather than resign. Sometimes we're in a village with philandering husbands, and whenever a wife learns that her own husband has been unfaithful she divorces (or kills, or castrates) him. There are also islands with blue- and brown-eyed inhabitants and wizards wearing hats. I'd be interested to learn of other variants and I'd like to track down the earliest statement of the problem.[2]

The well-known answer is that all ten mathematicians resign simultaneously on the tenth Friday following the super-mathematician's statement. To see this, start with fewer mathematicians. Suppose there are only two, A and B. Each thinks himself perfect and knows that the other has made a mistake.[3] So on the Friday following the statement of the super-mathematician, each expects the other to resign. A, for example, thinks "Since B knows that I haven't made a mistake, B must now know that he has made a mistake and will resign." But after that Friday, A realizes that B's failure to resign can only mean that B knows of a mistake made by A. So A resigns on the second Friday. B follows the same reasoning and resigns the same day. You can think through the case of three mathematicians to see that they all resign simultaneously on the third Friday. And with N mathematicians, each expects all the others to resign on the $N-1^{\text{st}}$ Friday. When that doesn't happen, each figures out that he himself must have made a mistake and resigns on Friday number N.

Now consider the following "solution" to the puzzle: "Obviously *nothing* will happen. The statement of the super-mathematician doesn't contain any new information—everyone already knows that someone has made a mistake—so after her statement nobody knows anything they didn't know before. Nothing has changed and life goes on."

This argument is incorrect. What, exactly, is wrong with it? That is, what is the fact that each mathematician knows only after the super-mathematician makes her statement, and not before?[4]

Here is an approximation to the answer: Once the super-mathematician speaks (but not before!), every mathematician knows that every mathematician knows that every mathematician knows that every mathematician knows that every mathematician knows that every mathematician knows that every mathematician knows that every mathematician knows that every mathematician knows that someone has made a mistake.

To be precise, let K_i^n denote the following statement: "For every sequence of n distinct mathematicians M_1, M_2, ..., M_n, it is the case that M_1 knows that M_2 knows that ... that M_n knows that i mathematicians have made a mistake." The actual fact we want is K_1^{10}, which is not exactly the long statement of the preceding paragraph. (The difference is that K_1^{10} applies only to 10-long chains of *distinct* mathematicians.)

To understand what's going on, let's use this notation to analyze the simpler case where there are only three mathematicians. Before the super-mathematician speaks, each mathematician believes that the other two have made a mistake; that is, K_2^1 holds. Furthermore, A thinks that B thinks that only C has made a mistake (remember, A thinks that B considers A error-free!). Each mathematician comes to the same conclusion about all the others, so we have K_1^2. Finally, A thinks that B would say that C knows of no errors at all; again, this is true for any triple of distinct mathematicians, and this is K_0^3. However, after the super-mathematician's statement, A knows that B knows that C knows of a mistake, and this is true of all triples—which is to say, K_1^3 is true.

Now we see the pattern. With ten mathematicians, we have initially K_9^1 and K_8^2 and K_7^3 and so forth up to K_0^{10}. Each time that the superscript increases by one, the subscript decreases by one, because each mathematician in the chain doesn't believe that he himself has made a mistake. All of these facts except the last are compatible with the super-mathematician's statement. But once everyone hears publicly that there is a mistake, K_1^{10} holds. The only way that K_1^{10} can fail to produce a resignation on the first Friday is if the tenth mathematician in each chain knows of an error that he did

not commit, which means the ninth mathematician in the chain now knows that there must be two errors; in symbols, K_2^9. Similarly we have K_3^8 after the second Friday, and so forth up to K_{10}^1 after the ninth Friday. That is to say, after nine Fridays each mathematician knows that ten mathematicians (not nine) have made mistakes, hence he himself has made one, and on the tenth Friday there is a sudden crowd at the unemployment office.

Super-mathematicians are expensive, so it's worth pointing out that we don't need one to set up the situation. Suppose on Monday one of the ten mathematicians stands up and says "I regret to announce that someone here has made a mistake!" (Any of them has the knowledge to do this, of course.) What happens? And as long as we're generalizing, what if k mathematicians all make this announcement simultaneously? What if one makes the announcement, and then immediately afterwards $k-1$ chime in with agreement? What if one mathematician asserts that m of the mathematicians have made errors? In the original puzzle, what if only k of the mathematicians (not 10) have actually made a mistake? What if, after two Fridays, one of the non-super-mathematicians announces that at least six of the others have made errors? These are easy problems if you've read this far. I think the first of these variants, that is, $k = m = 1$ and no super-mathematician, makes for a cleaner version of the puzzle than the original.

We find something amusing by looking harder at the wildly artificial conditions of the problem. Clearly it's crucial that each mathematician has made a mistake, that at least one mistake of each mathematician has been detected, that everyone knows of everyone else's mistakes, and so forth, otherwise there can be holes in the reasoning (e.g., A, knowing of a mistake in B's work, would have to consider the case that C did not *yet* know of B's mistake). But even so we haven't said enough. It does not suffice for everyone to know of everyone else's mistakes. We also must assume that everyone knows *that everyone knows* of everyone else's mistakes, and so forth up to a chain of ten. Furthermore, it's not enough to assume that each mathematician is a perfect reasoner. Each must know that the others are all perfect reasoners, and that all the others know this as well, and that everyone knows that everyone knows that everyone is a perfect reasoner, and so forth. But of course we can't say all this explicitly in the problem statement because it gives too much away.

Now let's take a closer look at the "knows that" construction. As we've seen, the statements "F is true", "A knows that F is true", "B knows that A knows that F is true" and so forth are completely distinct and can be true or false independently. This is true even when the chain contains duplicates; it's possible that A knows that B knows F, but B doesn't know that A knows that B knows F. The important point is that a public announcement, as in this problem, instantly extends the chain indefinitely. Also, this "public knowledge" effect is the reason that we don't have to state explicitly the additional assumptions of the preceding paragraph; there's an implied *public* statement that all mathematicians are perfect reasoners, which suffices.

A more complicated chain of what-does-someone-know predicates is illustrated by this puzzle[5]:

> Two positive integers are chosen. The sum is revealed to logician A and the sum of the squares to logician B. Both A and B are (publicly!) given this information and the information contained in this sentence. The conversation between A and B goes as follows, with B starting:
>
> B: "I can't tell what the two numbers are."
>
> A: "I can't tell what the two numbers are."
>
> B: "I can't tell what the two numbers are."
>
> A: "I can't tell what the two numbers are."
>
> B: "I can't tell what the two numbers are."
>
> A: "I can't tell what the two numbers are."
>
> B: "Now I can tell what the two numbers are."
>
> What are the two numbers?

Here, it's not just "A knows that B knows that . . . " but something more involved.

Notes

[1] Concerning gender: We assume that all mathematicians are male and all super-mathematicians are female. Apologies if this approach to the pronoun problem is not to your taste.

[2] The version we're using has a flaw: If the mathematicians are such perfect reasoners, how is it that they make mistakes in their work? We could avoid this infelicity by switching to philandering husbands—there's no conflict between adultery and logic, and philandering husbands are more plentiful than mathematicians anyway. Nevertheless, I have to stick with the mathematicians, being myself completely unfamiliar with infidelity. So for us the term "perfect reasoner" really means "perfect reasoner except for making an occasional mistake in his own work." The perfect reasoner assumption, by the way, is widespread in puzzles of this sort, applied equally to logicians, pirates with coins, knights and knaves, and many more.

[3] Is it possible to "know" something that's actually false? This is a deep problem of philosophy, but for our purposes it's only a question of English usage. We really don't care about any distinction between "A thinks F" and "A believes F" and "A knows F". When it's known that A has made an error, I write "A thinks that his work is error-free" instead of "A knows that his work is error-free" only because the latter sounds funny.

[4] I first saw this followup in Spivak's *Calculus* (Fourth Edition) where it's given as a problem (in different words). Probably it too goes way back.

[5] Tom Ferguson, Mathematics Magazine, May 1984, p 180 (thanks to David Cantor, UCLA).

Acknowledgements. Thanks to Steve Denenberg, John Draper, Pattabhiraman Krishna, Anirudh Suresh, and Bart Wright for helpful discussion.