'To Queue Or Not To Queue'

That is the question not only for millions of consumers but for Israeli researchers, who have assumed a prominent position in the field of queuing theory

Ziv Hellman

ONE OF THE FIRST – AND most annoying – aspects of Israel that many newcomers from English-speaking and other more refined countries notice is how rare it is for the locals to stand politely in an orderly line. Whether getting on a bus or buying ice cream, forming a straight line seems like the most foreign idea to Israelis, who instead clump into unruly masses.

Post offices, health clinics and public offices try to combat this tendency by issuing numbered tickets to customers in the order they arrive. But even then citizens seem to find ways to avoid simply waiting their turn, such as taking several tickets, going out to run errands and returning just in time to be served. To an observer used to well-ordered lines, it can appear to be a mess.

Given this reality, it might seem anomalous that the field of queuing theory, a mathematical discipline that studies the subject of lines and waiting in lines, counts several Israelis among its founders and prominent practitioners. On second thought, perhaps it makes perfect sense, when you consider that Israeli researchers, such as Refael Hassin of Tel Aviv University, who has composed an entire book (along with his colleague Moshe Haviv of the Hebrew University) entitled "To Queue or Not to Queue," have also been at the forefront of research indicating that the orderly queue so beloved in North America and Britain may not be the best way to serve customers – and that just the opposite is true.

"There is no doubt about it," says Hassin. "The orderly, first-in, first-out line is the least efficient system of them all, from the perspective of social optimality" or, in other words, the sum of what each person standing in line gains and loses out of the process. Israelis, it seems, may have had it right all along, actually increasing efficiency in their disorganized queues.

Queuing theory is a specialty in operations research, which itself is an applied branch of mathematics, studying optimal or near-optimal solutions to complex problems. It started in 1909, when a Danish engineer working for the Copenhagen Telephone Exchange conducted the first rigorous study of waiting in lines. It has since then found major application in the fields of transport, telecommunications, call-centers, traffic flow and network routing. The explosion of Internet usage over the past 15 years has spurred interest in the study of reducing waiting times for computer servers.

But the earliest article to study individual behavior in queuing from the perspective of social optimality was published by a Technion researcher, Pinhas Naor, in 1969. (Tragically, Naor was killed in an airplane crash shortly afterwards, in December 1970, while en route to Europe to organize an international conference on Operations Research.) Calculating social optimality entails taking into account that each individual gains something from the service received at the end of the wait in line, but pays a price in the time sacrificed during the wait – although different people may have differing assessments of both the benefit and the cost. Society has an interest in maximizing the sum total of individual benefits and minimizing the costs, which researchers like Naor and Hassin attempt to assess.

ARIEL JEROZOLIMSKI

NON-QUEUE: Swarming onto a bus in Jerusalem

IN NAOR'S ANALYSIS, THE PROBlem with the first-come, first-served system is that no individual considers the costs to anyone else from that fact that he holds his position in a queue. "People in lines tend to think only about themselves and ignore their impact on others," explains Hassin. "If I join the line and you come later, you will wait longer because of me. This leads to selfish behavior, ignoring the effect our actions have on others."

Naor's proposed solution to this problem was levying a charge on individuals wishing to join a queue, with the cost tied to the longer wait each person in the queue imposes on others. In theory, just as imposing fees or fines on polluters, because they benefit at the expense of others, reduces pollution, charging people for the right to enter a line would result in socially more optimal lines.

Hassin notes that standard economics models indeed indicate that optimality can usually be attained when all relevant costs are borne by those benefitting from a given situation. But calculating the optimal price to charge requires a detailed model taking into account each individual's costs and benefits – information that most service providers cannot hope to gain about all their potential customers. There is the additional challenge of ensuring that potential customers tell the truth regarding costs and benefits.

Pondering this matter led Hassin to study the possibility of attaining more socially beneficial results by looking at alternatives to the first-come first-served model that is nearly universally regarded as the correct

one to strive for. He has intriguing arguments showing that the traditional queue has significant disadvantages. Suppose, he says, you are at a party at which beers are being handed out, one per customer, at two opposite corners of a room. In one corner, an orderly queue is formed, while at the other one the beer is handed out according to a random lottery. "If you think about it for a while, you come to the conclusion that you would expect two-thirds of the people at the party to join the random queue, instead of the orderly line," says Hassin. "Imagine seeing the orderly line filled with at least one-third of the guests. If you go to the end of that line, you know you've got quite a wait ahead of you. With the random one, you've got a chance to do better. In that situation, most people understand that a random ordering can be preferred to an ordered one."

HIS RESEARCH INTO THE SUBJECT led him almost literally to "turn the problem on its head" and recommend a "reverse queuing" system in which queues "run backwards," with the last customer entering being the first to be served. His first major publication on the subject, in the prestigious journal "Econometrica," in 1985, attracted attention for his surprisingly counter-intuitive conclusion that a reverse queue would be the best possible system.

In this so-called LIFO (last in, first out) system, the main question becomes how determined a person who is already in line is to "stick it out" as the line in front of him grows, or to opt out. Presumably, a person would opt out when the expected cost of waiting is greater than the benefit of waiting – which is what people consider every time they join a line, the difference being that in the LIFO system, an individual opting out of the line affects only himself in his actions, no one else. The addition of a new person in line would impose no costs at all on those who come later. Lines would in general be shorter and, on average, we'd spend less time waiting.

The literature explaining Hassin's result sometimes uses the stylized example of a water fountain in a city park, with a line of thirsty individuals waiting patiently for their turn. Ideally, the best system would enable the water fountain to be in maximal use, slaking people of their thirsts, with as few people as possible at any given moment waiting around, wasting their time.

$$n^* = \lfloor \nu^* \rfloor \leq \lfloor g(\nu^*) \rfloor = \left\lfloor \frac{R\mu}{C} \right\rfloor = n_e.$$

MATH QUEUE: An equation from Hassin's publications, calculating the maximal number of customers in a queue that is consistent with social optimality

Presumably there is a maximum line length people are willing to tolerate before not even trying to stand in line, of, say, 14 or 15 people. When the line is that long, people pass it by. If it falls below that threshold, someone always joins at the end to bring it back to that length. In the first-come first-served system, then, one gets the worst possible result – the line is always at the maximal length.

If, instead, each new arrival goes to the front of the line, one would expect that individuals would only enter the line when it is short enough for them to assume that they have a good chance of remaining at the front, given that other people may be joining the line. If not, they would give up and opt out of the line long before it reaches its maximal length. The result, it turns out, is shorter lines overall, with people spending less time just waiting. Since under both the standard queue and the LIFO system the water fountain will continuously be in use, water consumption is the same in both cases, but because lines are generally shorter under LIFO, waiting time is reduced, which is why this solution is optimal. The model aligns personal incentives with "socially optimal goals."

> Some researchers are confident that businesses adopting queuing systems that differ from the traditional first-come, first-served model will actually see an increase in their bottom lines

One problem with this model is that people located at the end of the line may leave it and immediately return, thereby being placed first, a situation that could be rectified by keeping tabs on who left and not letting him or her return immediately afterwards. Another objection to the model is that it ignores the possible "risk aversion" of customers, because a customer who is currently last in line continually runs the risk that the line will grow indefinitely longer, pushing him farther and farther away from service. But Hassin has a response to the objections with a modification of the model in which newly arrived customers are placed directly in front of the last person on line, so that everyone who is not last runs no risks, and only the person last in line needs to run cost-benefit analyses.

It is difficult to imagine any public office or business adopting a LIFO customer service or even one that officially lets newcomers jump the queue ahead of only the last person in line, but Hassin is confident that businesses adopting queuing systems that differ from the traditional first-come, first-served model will actually see an increase in their bottom lines. "The first-come, first-served model is the worst possible one, from the perspective of social optimality," he tells The Report. "Even conducting a lottery among all the people waiting for service at any given time to determine who will get served first is better."

Take for example, customers at a bakery. Customers who crowd the sales clerks – as often happens in Israel – instead of patiently waiting in an orderly line, will often experience shorter waiting times, so that more sales can be made and more money can fill the till.

"If there are 10 people in the shop and there is no organized line, on average any one will be served after the fifth person," says Hassin. "Of course, someone will get served 1st and 2nd, and someone will have to be served last. But consider it from the point of view of human decision-making strategies: If one is deciding whether or not to enter a shop and sees many people there already, most would prefer an unordered queue – because then there is a good chance of being served sooner than if one were

waiting patiently in line. And that could attract more customers."

HASSIN, 58, A PROFESSOR AT TEL Aviv University's Department of Statistics, received a BA in economics from Tel Aviv University in 1973, and went on to earn a PhD in Operations Research at Yale University in Connecticut. In addition to teaching at Tel Aviv University, he has also held positions at Stanford University, IBM's Research Division, George Washington University and Carnegie-Mellon. A prolific researcher, he has published over 100 journal articles.

In much of the analysis of social aspects of operational research and economics, efficiency often seems pitted against fairness. Hassin is well aware that moving away from the standard first-come, first-served paradigm is likely to strike most people as unfair, and he notes that some of his colleagues are now studying queuing theory from an axiomatic perspective of fairness, but that ultimately, in his opinion, it comes down to psychology, while he is concentrating on efficiency. He would also argue that, in the broadest perspective, striving for efficiency proves to be better for all involved.

For example, it might seem fair to set a uniform and low price for tickets to a major event, giving everyone an equal opportunity to stand in line and obtain a ticket, until availability runs out. But that "one price for all" may be an illusion, because waiting in line is a cost in itself, so that in effect those who are willing to "pay more" by standing in line end up getting the ticket ahead of those who are willing to pay less out of their time budget. For many, it seems wrong in that situation to allow people to pay others to stand in line for them, but Hassin sees it as a standard economic exchange – money for someone else's time – and it is especially efficient if that transaction frees the payer to do something more constructive during the time he would otherwise be standing in line. "The person paying and the person receiving money to stand in line both do so because they feel they would be better off," he argues. "So why not let both of them do what's best for them?"

HASSIN'S ENDLESS CURIOSITY regarding queuing has led him to study many different questions relating to the topic in his career. For example, there are many places in which people begin queuing for items even before shops or offices open to give themselves as short a waiting time in the morning as possible. Hassin wondered whether total waiting times could be reduced by conducting a lottery among everyone who shows up at the opening hour, thus rendering pre-opening lines useless. But a careful study of the issue conducted by him and his former student Yana Kleiner showed that conducting such a system would achieve nothing in terms of reducing waiting times – because it would give an incentive to people who might otherwise show up later to arrive at opening time, knowing that doing so would give them an opportunity to be first in line. "That analysis indicates that it might be best to conduct a lottery among everyone who shows up, at repeated intervals several times a day," says Hassin. "That way arrivals can be spaced out and total waiting times reduced on average."

> 'Israelis, it seems, may have had it right all along, actually increasing efficiency in their disorganized queues'

Another subject he looked into is the question of the best placement of a petrol station along a stretch of motorway. Suppose there are two such stations, one after the other. Motorists have the option of entering the first station, or passing it by if the line appears too long there (and they have sufficient fuel to get to the next station), but by doing so they are committing themselves to entering the second station, no matter how long its line is. What would people do in such a situation? "That question exercised my mind for a while," admits Hassin. He finally came to the conclusion, complete with a mathematical proof to support it, that the owner of the first station has an advantage. "People's behavior is characterized by a 'threshold' length – drivers keep in mind a maximal line length they are willing to tolerate," says Hassin, "and tend to choose the first line if it is shorter than this threshold. This results overall in more drivers [entering] the first station, so the first station wins."

With regard to the perennial question of whether, say, a bank should have several parallel lines, one for each teller, or one line directing customers to the next available teller, Hassin has no hesitancy in always recommending the one-line approach, as queuing theory indicates the multiple lines method is hopelessly sub-optimal with respect to average waiting times. "With one line, you also avoid situations in which a server is under-utilized, while others are over-subscribed," he adds. "The work load is then more optimally spread."

Hassin has also been studying questions of whether it is to customers' detriment to give them too much information regarding queues. "Because customers in standard queues ignore the costs they impose on others," he says, "a better social outcome can often be obtained by concealing the lengths of queues so they act according to averages and not the actual length of the queue."

The matter of how much information may be "too much for your own good" is closely related to a phenomenon called Braess's Paradox, after the German mathematician Dietrich Braess, in which, surprisingly, adding more roads connecting two cities can actually lead to greater traffic jams and longer commuting times. The paradox is not an obscure theoretical possibility – there are examples from real life in which the paradox's bite has been felt keenly by commuters, as when the closure of newly constructed roads in Stuttgart, Germany, for servicing improved the flow of traffic, and in New York City, where contrary to dire predictions, the closure of 42nd Street to vehicular traffic has actually reduced congesting the area. As Hassin explains, Braess's Paradox essentially follows from people switching to new possible routes once they are made available or known to them – routes that are better to each individual if only a minority were to switch, but are worse for everyone if many people switch at once. The same phenomenon can occur, it turns out, when considering queues.

Finally, reflecting on the behavior of his fellow countrymen with respect to polite queuing, Hassin recommends not judging them too harshly. "Israelis behave in a more strategic and less naive way [than people in many other countries]," he notes. "They calculate how to shorten waiting times. For example, when coming to a stop in front of traffic lights, Israelis choose their lanes to bring all the lines to approximately the same length, which is not something typically seen in the United States, for example. Israelis behave in a way that is closer to what the mathematical models of rationality predict."